ESTIMATION DE LA VALEUR D'UN PORTEFEUILLE D'EPARGNE AVEC DES TECHNIQUES DE MACHINE LEARNING

Manuella MOUAFO

Remerciements

Je tiens à remercier toutes les personnes qui ont contribué à la réalisation de ce mémoire.

Je tiens tout d'abord à remercier le Groupe CNP Assurances de m'avoir accueillie au sein de ses différentes équipes afin d'y effectuer mon alternance pour ma dernière année d'étude en actuariat.

Je souhaite ensuite remercier Hamdi KACEM, mon tuteur professionnel, pour son implication, sa disponibilité ainsi que l'ensemble de ses conseils avisés tout au long de la réalisation et de la rédaction de ce mémoire.

Je remercie tous les membres de la Direction Risque Groupe pour leurs encouragements, et leur accompagnement, en particulier Yann DESSEPRIT, Mylène CHEVALIER, Alaeddine FALEH et Tarek AOUDI.

Je tiens également à remercier la Direction Technique et de l'Innovation Groupe de m'avoir accueilli pendant ma période de stage initiale.

j'adresse mes remerciements à tout le corps professoral de l'ISUP et notamment à Charlotte DION, ma tutrice académique.

Enfin, je souhaite remercier ma famille pour m'avoir soutenue tout au long de ma scolarité.

Résumé

Afin d'évaluer la rentabilité ajustée au risque, les assureurs développent des modèles Actif/Passif de plus en plus sophistiqués et complexes. Dans ce mémoire, nous explorerons l'utilisation des techniques de Machine Learning pour estimer la valeur d'un portefeuille d'épargne (que nous noterons Value In Force ou VIF) en se basant sur des données et en identifiant les facteurs de risque clés qui influencent les résultats. Nous proposons une approche qui permet de reproduire dans une certaine mesure le comportement du modèle Actif/Passif. Les résultats de cette étude ont pour but de fournir un dispositif rapide et fiable d'aide à la décision.

Les méthodes de Machine Learning permettent d'apprendre une variable à expliquer à partir de variables explicatives sans être explicitement programmées. Nous avons développé des modèles pour apprendre la VIF future en fonction de l'environnement économique, et de l'ensemble des hypothèses clés en entrée des modèles actuariels de projection Actif/Passif.

Les modèles retenus sont le XGBoost, le random forest, et la régression LASSO. Les deux premiers ont été choisis pour leurs réputation dans le domaine et leur aptitude à capter des phénomènes complexes. La régression LASSO a été choisie afin de proposer un moyen simple et facilement compréhensible malgré l'hypothèse de l'existence d'une relation linéaire entre les variables explicatives et la variable à expliquer.
Nous avons aussi proposé une approche par décomposition en sous variables, qui a pour but de prédire individuellement les trois principales variables qui constituent la VIF.

Il en ressort que la performance des méthodes retenues est fortement dépendante de la qualité de la base d'apprentissage. En effet, l'enrichissement de la base demeure un levier important permettant l'amélioration des résultats. Le modèle XGBoost avec une approche par décomposition en sous variables présente les meilleures performances et des résultats satisfaisants.
Nous avons proposé une application concrète de ce modèle à travers le calcul du SCR.
Nous avons d'autre part utilisé un module d'explicabilité des modèles de Machine Learning pour apprécier la cohérence des traitements restitués par notre modèle.

Notre approche exploratoire nous a induit qu'il reste plusieurs pistes d'investigations à entrevoir pour améliorer les résultats. En effet, ce genre de modélisation n'est pas sans risque d'erreurs, avec des biais qui peuvent être de sources différentes. On note par exemple le risque d'avis d'expert, de sélection de données et de de variables omises.

Abstract

In order to assess risk-adjusted profitability, insurers are developing increasingly sophisticated and complex asset/liability models. In this dissertation, we will explore the use of Machine Learning techniques to estimate the value of a savings portfolio (which will be called Value In Force) based on data and identify key risk factors that influence the results. We propose an approach that can replicate to some extent the behavior of the Asset/Liability model. The results of this study are intended to provide a fast and reliable decision support tool.

Machine Learning methods allow to learn a variable to be explained from explanatory variables without being explicitly programmed. We have developed models to learn the future VIF according to the economic environment and the set of key input assumptions of actuarial asset/liability projection models.

The models selected are the XGBoost, the random forest, and the LASSO regression. The first two were chosen for their reputation in the field and their ability to capture complex phenomena. The LASSO regression was chosen in order to propose a simple and easily understandable method despite the hypothesis of the existence of a linear relationship between the explanatory variables and the variable to be explained.
We proposed a direct approach which consists in directly predicting the VIF and an approach by decomposition in sub-variables, which aims at predicting individually the three main variables which constitute the VIF.

It appears that the performance of the selected methods is strongly dependent on the quality of the learning base, in fact the enrichment of the base remains an important lever allowing the improvement of the results. The XGBoost model with an approach by decomposition into sub-variables presents the best permanence and satisfactory results.
We have proposed a concrete application of this model through the calculation of the SCR.
We have also used an explicability module for Machine Learning models to assess the coherence of the treatments returned by our model.

Our exploratory approach has led us to believe that there are still several avenues of investigation to be explored to improve the results. Indeed, this type of modeling is not without risk of error, with biases that can come from different sources. For example, there is the risk of expert opinion, selection, data and omitted variables.

Note de synthèse

L'assurance est un secteur clé qui a un impact significatif sur l'économie mondiale en protégeant les individus et les entreprises contre les pertes financières imprévues. La prédiction précise et fiable des résultats financiers d'une entreprise d'assurance est donc primordiale pour assurer sa stabilité financière et sa durabilité à long terme.

Afin d'évaluer la rentabilité ajustée au risque, les assureurs développent des modèles Actif/Passif de plus en plus sophistiqués et complexes. Dans ce mémoire, nous explorerons l'utilisation des techniques de Machine Learning pour estimer la valeur d'un portefeuille d'épargne (que nous noterons Value In Force) en se basant sur des données et en identifiant les facteurs de risque clés qui influencent les résultats. Nous proposons une approche qui permet de reproduire dans une certaine mesure le comportement du modèle Actif/Passif. Les résultats de cette étude ont pour but de fournir un dispositif rapide et fiable d'aide à la décision.

La VIF correspond à la valeur actuelle probable des résultats futurs, nets d'impôts, générés par le portefeuille de contrats en cours . Elle représente la quantité de richesse générée dans le futur à partir des contrats en cours au sein du portefeuille et revenant à l'assureur.
La VIF est calculée à l'aide de modèles de projection des flux de trésorerie futurs et permettant de relier la dynamique du passif à celle de l'actif et notamment l'application de la stratégie d'investissement et de revalorisation.

En effet, les modèles actuariels Actif/Passif reposent sur l'exploration d'un nombre élevé de scénarios via les approches stochastiques avec plusieurs simulations afin d'évaluer leurs engagements. Ces modèles nécessitent ainsi en entrée un ensemble d'informations riche et varié tant sur l'environnement économique, tant sur la composition d'actifs et de passif mais aussi des hypothèses et des décisions de management rendant les calculs complexes.

Le calcul de la rentabilité d'un produit se fait en évaluant les résultats futurs probables générés par la commercialisation de ce produit. La projection d'un compte de résultat est donc réalisée pour chaque année à venir jusqu'à la fin des engagements pris par l'assureur ou alors sur une période suffisamment longue à l'issu de laquelle les engagements restant son négligeables. La période de projection utilisée dans notre cas est de 50 ans.
Il s'agit d'un processus long et chronophage, d'où notre volonté de mettre en place une alternative fiable et rapide, capable de servir comme dispositif de pilotage de sensibilités et d'aide à la décision, à l'aide du Machine Learning.

Le Machine Learning est une technologie de l'intelligence artificielle permettant à des algorithmes d'apprendre sans avoir été explicitement programmées à cet effet. Les algorithmes de Machine Learning, plus précisément d'apprentissage supervisé seront donc utilisés dans une démarche prédictive dans le but d'estimer la VIF. Nous fournirons à l'algorithmes les données et hypothèses du modèle Actif/Passif et notre modèle de Machine Learning devra essayer de reproduire ce dernier et de restituer la VIF.
Le schéma ci-dessous résume notre approche de façon simple :

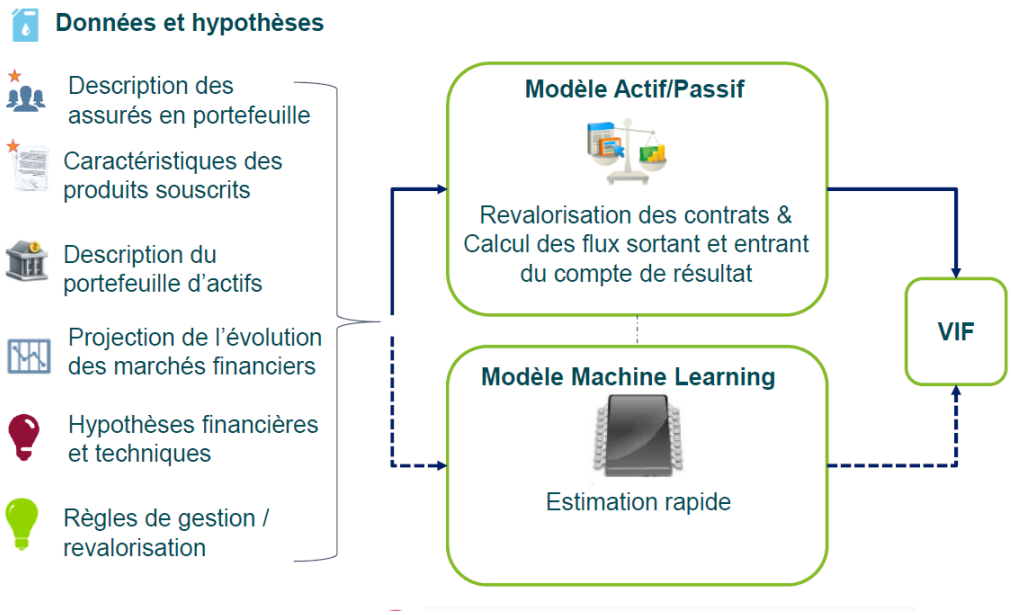

FIGURE 1 – Approche Machine Learning

La mise en place en place d'un modèle de Machine Learning repose sur 4 principales étapes :

— La construction de la base de données
— La sélection des modèles et des techniques de scoring
— La sélection des variables
— L'amélioration et la validation du modèle

Nous avons construit notre base de données en nous basant sur les simulations effectuées par l'outil de modélisation interne de CNP Assurances. Nous avons incorporé à notre base de données les éléments pris en in-put dans le modèle et la VIF ressortie.
Nous avons fait le choix de constituer la base d'apprentissage en effectuant plusieurs jeux de sensibilité touchant à la fois les hypothèses économiques, techniques et financières. L'objectif étant de constituer une base fiable permettant de capter la sensibilité et le comportement du modèle Actif/Passif dans différents environnements et configurations, nous considérons les observations relatives à chaque scénarios stochastiques.
Cette base est répartie en 3 catégories : les données de l'environnement économiques issues du GSE, les données contractuelles (relatives au contrat signé par l'assuré) et du passif, les données visant la composition du portefeuille d'actifs.

Les modèles retenus pour aborder ce problème sont le XGBoost, le random forest, et la régression LASSO. Les deux premiers ont été choisis pour leurs réputation dans le domaine et leur aptitude

à capter des phénomènes complexes. La régression LASSO a été choisie afin de proposer un moyen simple et facilement compréhensible, malgré l'hypothèse de l'existence d'une relation linéaire entre les variables explicatives et la variable à expliquer.

Nous avons proposé une approche directe qui consiste à prédire directement la VIF et une approche par décomposition en sous variables, qui consiste à prédire individuellement les trois principales variables qui constituent la VIF puis à les agréger. En effet, la VIF se décompose comme suit :

$$VIF = Prélèvements - Chargements - Coût$$

Avant de mettre en place nos modèles, nous avons procédé à une sélection des variables. Cette étape consiste à choisir les variables les plus pertinentes pour le modèle final et permet de réduire la complexité du modèle et améliorer l'interprétabilité. Pour le faire, nous avons dans un premier temps observé les corrélations entre nos variables comme le montre le graphe ci-dessous :

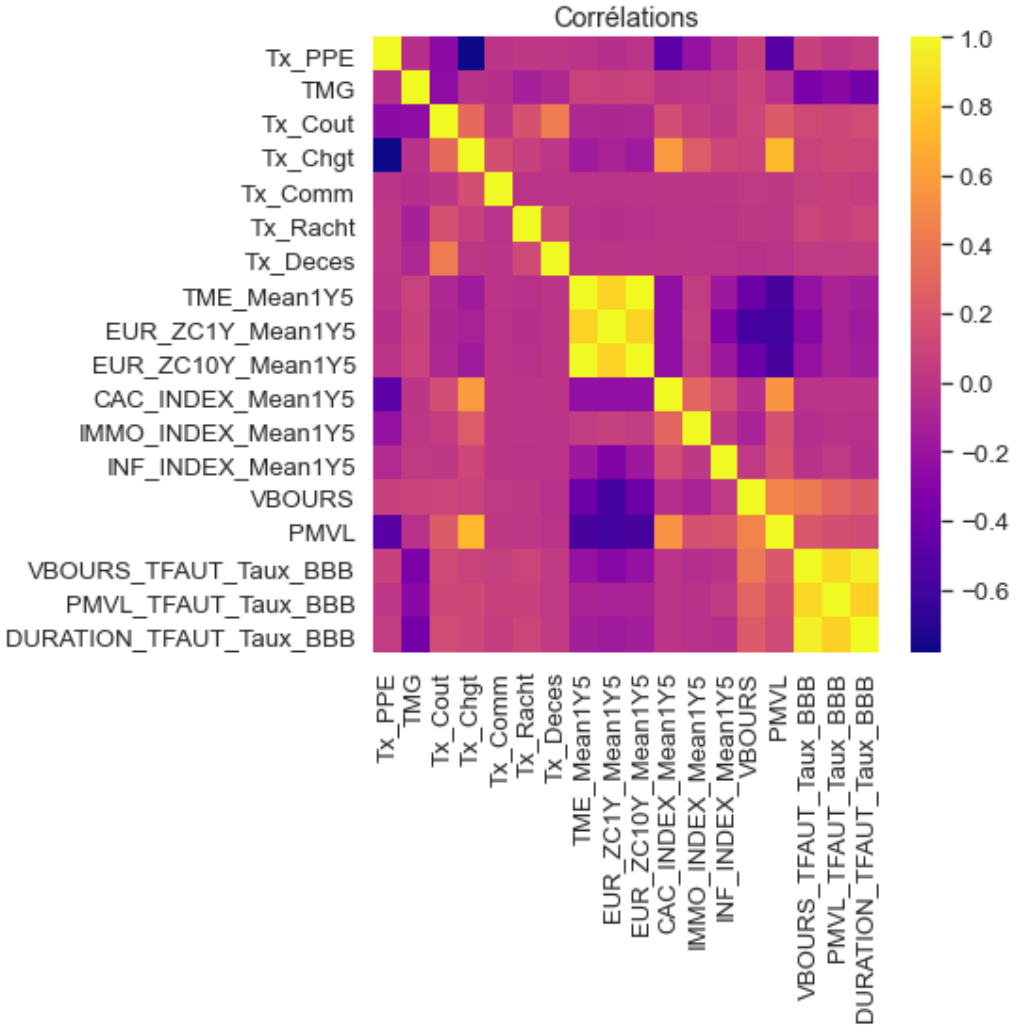

FIGURE 2 – Corrélation des variables explicatives

L'observation des corrélations montre une forte corrélation entre le TME et les zéro-coupon ; entre la duration, la valeur boursière (VBOURS) et les plus ou moins value latentes (PMVL) par groupe d'actifs. Ensuite, le module feature importance des modèles XGBoost et random forest nous ont permis d'observer les variables qui ont été les plus déterminantes. Les coefficients LASSO nous ont également donné une vision sur les variables qui avaient été utilisées pour le modèle ou pas. Nous avons donc

décidé de ne garder que le TME et la duration pour la suite de l'exercice.

Une fois la sélection des observations effectuées, nous avons pu mettre en place nos différents modèles. Nous optimisons les paramètres de chaque modèle en utilisant la méthode de validation croisée permettant de choisir les paramètres a priori les plus optimaux pour chaque modèle.

Ci-dessous les premiers résultats à la maille globale :

Modèle	R2	RMSE (En M)	MRE	MAE (En M)
LASSO	0,75	1 850	180%	1423
RF	0,92	545	47%	238
XGB	0,99	289	15,7%	85

TABLE 1 – Résultats à une maille globale I

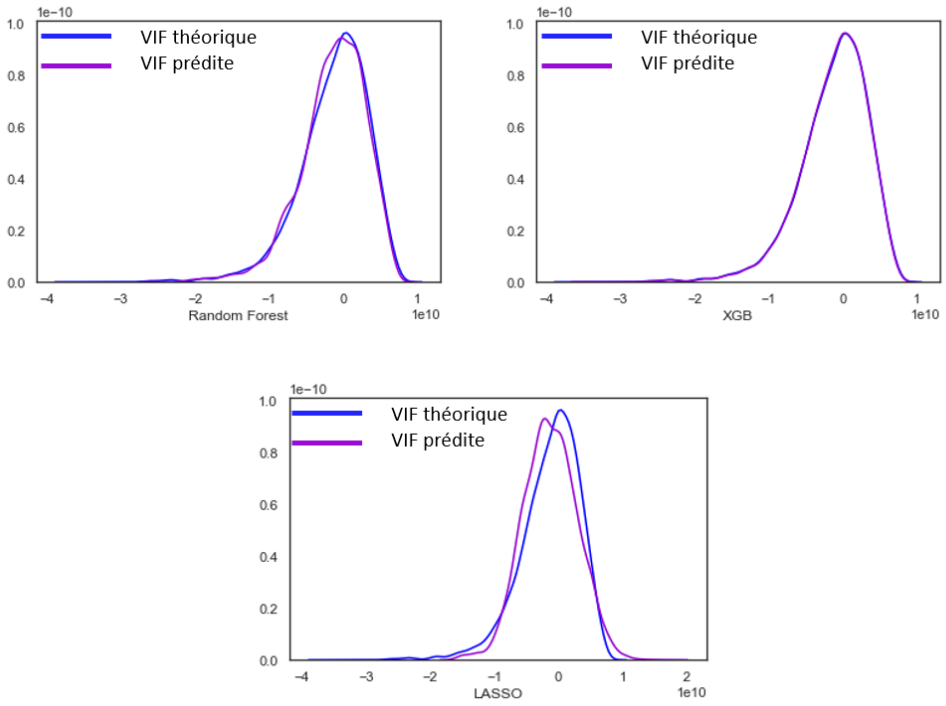

FIGURE 3 – Comparaison des densités I

Ces résultats nous montrent que le modèle XGBoost est meilleur que les autres, avec une densité qui se rapproche parfaitement de la densité théorique de la VIF. Il possède un R2 très élevé de 0,99, donc le modèle parvient à correctement deviner la distribution de notre variable. Son erreur relative absolue moyenne de 15,7% signifie que pour chaque observation prédite, le modèle commet une erreur d'environ 15,7% en plus ou en moins sur l'Equity. Cette déviation se retranscrit pas un écart de 85M € plus ou moins.

Nous avons jugé ces résultats peu satisfaisants et avons entrepris d'améliorer nos résultats en effectuant de nombreuses sensibilités, permettant de capter le comportement de notre modèle Actif / Passif

dans différentes configuration impacts certaines hypothèses clés comme le niveau des chargements, de commissions et de coûts...
On peut observer ci-dessous les nouveaux résultats qui se sont montrés plus concluants :

Modèle	R2	RMSE (en M)	MRE	MAE (En M)
LASSO	0,83	1 702	180%	1 337
RF	0,99	83	9,7%	58
XGB	0,99	75	7,2%	48

Table 2 – Résultats à une maille globale II

Les résultats peuvent aussi être observés à une maille plus fine, par type de chocs sur le Closing :

TypeChocSCR	R2 LASSO	R2 RF	R2 XGB	RE LASSO	RE RF	RE XGB
Central	0,83	0,99	0,99	6,0%	2,5%	2,3%
ChocActionType1	0,79	0,99	0,99	-10,3%	-0,2%	0,0%
ChocActionType2	0,84	0,99	0,99	-10,4%	0,1%	0,1%
ChocBaisseRachat	0,50	0,99	0,99	10,0%	3,1%	2,9%
ChocBaisseTaux	0,79	0,99	0,99	-8,1%	2,5%	-2,7%
ChocFrais	0,81	0,99	0,99	-10,2%	-0,8%	-0,2%
ChocHausseRachat	0,87	0,99	0,99	1,7%	0,0%	-0,1%
ChocHausseTaux	0,77	0,99	0,99	-3,6%	-0,2%	-0,2%
ChocImmobilier	0,81	0,99	0,99	-13,0%	-0,6%	-0,6%
ChocLongevite	0,80	0,99	0,99	17,2%	0,0%	-0,4%
ChocMortalite	0,83	0,99	0,99	8,6%	0,3%	0,0%

Table 3 – Résulats à une maille Closing et par choc II

On constate que chacun de nos modèles s'est amélioré, en particulier le random forest qui est passé d'un RMSE de 545 M à une RMSE de 83 M. Le XGBoost reste le meilleur, avec une RMSE de 75M et une MRE de 7,2%. Ses erreurs relatives sur les différents types de choc sont de façon globale, inférieures à celles du Random Forest.
Nous avons ensuite utilisé sur notre modèle le plus performant (XGBoost) une méthode de sélection des observations qui nous a permis de réduire l'erreur du modèle. On a observé par exemple que l'erreur relative sur le Closing central est passée de 2,3% à 1,6%. Nous avons utilisé l'approche par décomposition en sous variables, qui nous a fourni les résultats suivants :

TypeChocSCR	R2 XGB	R2 ssVar	RE XGB	RE SS VAR
Central	0,99	0,99	1,6%	1,0%

Table 4 – Résulats à une maille Closing et par choc IV

Modèle	R2	RMSE (en M)	MRE	MAE (en M)
VIF XGB	0,99	34,8	2,6%	20,3
VIF Ss variables	0,99	28,0	1,4%	15,4
Coûts	0,99	5,7	0,2%	3,3
Commissions	0,99	8,5	0,6%	4,6
Prélèvements	0,99	13,4	0,7%	7,9

TABLE 5 – Résultats à une maille globale IV

Le fait de séparer notre variable cible en 3 variables moins complexes nous a permis d'améliorer légèrement notre score, passant d'une erreur relative de 1,3% à 1,0% sur la VIF centrale. On peut observer que le modèle arrive à capter parfaitement la densité de la variable théorique et le R2(0,99) montre que le modèle comprend bien la complexité cachée dans le calcul de la VIF.
La méthode par sous variable avec le modèle XGBoost est donc celle qui offre de meilleurs résultats.

Nous avons appliqué l'algorithme de transparence SHAP à notre modèle afin de mieux apprécier les résultats. Ce module nous a permis d'obtenir des informations complémentaires quant aux règles de décisions du modèle et la façon dont la VIF est estimée.

Afin de mettre en application nos travaux, nous avons calculé les SCR marché et souscription vie. Ci-dessous les résultats obtenus :

SCR calculé	Valeur théorique (en M)	Valeur prédite (en M)	Ecart (en M)	Erreur relative
SCR marché	60,7	60,0	0,69	1,1%
SCR vie	4 076	4 113	-37,23	-0,9%

TABLE 6 – Résultats du calcul des SCR par module

On observe des erreurs de 1,1% et -0,9%, ce qui est un résultat plutôt satisfaisant. Le modèle que nous avons mis en place nous paraît relativement fiable en matière de résultats.

Notre approche exploratoire nous a convaincu qu'il reste plusieurs pistes d'investigations à entrevoir pour améliorer les résultats. En effet, ce genre de modélisation n'est pas sans risque d'erreurs, avec des biais qui peuvent être de sources différentes on note par exemple le risque d'avis d'expert, de sélection, de données et de variables omises.
La base d'apprentissage pourrait être enrichie à l'aide de nouvelles sensibilités et on pourrait éventuellement trouver de meilleurs méthodes pour incorporer la stratégie d'investissement et les règles de gestion et de revalorisation spécifiques au portefeuille étudié.
Il pourrait être envisageable de tester d'autres modèles de Machine Learning tels que les réseaux de neurones afin d'observer leur apport.

Summary

Insurance is a key industry that has a significant impact on the global economy by protecting individuals and businesses from unexpected financial losses. Accurate and reliable prediction of an insurance company's financial results is therefore critical to its long-term financial stability and sustainability.

In order to assess risk-adjusted profitability, insurers are developing increasingly sophisticated and complex asset/liability models. In this dissertation, we will explore the use of Machine Learning techniques to estimate the value of a savings portfolio (which we will note Value In Force) based on data and identify key risk factors that influence the results. We propose an approach that can replicate to some extent the behavior of the Asset/Liability model.
The results of this study are intended to provide a quick and reliable decision support tool.

The VIF corresponds to the probable present value of future results, net of taxes, generated by the portfolio of current contracts. It represents the quantity of wealth generated in the future from the contracts in force within the portfolio and accruing to the insurer.
The VIF is calculated using models that project future cash flows and make it possible to link the dynamics of the liabilities to those of the assets and, in particular, the application of the investment and revaluation strategy.

Actuarial Asset/Liability models are based on the exploration of a large number of scenarios via stochastic approaches with several simulations in order to evaluate their commitments. These models therefore require a rich and varied set of input information on both the economic environment and the composition of assets and liabilities, as well as assumptions and management decisions that make the calculations complex.

The profitability of a product is calculated by evaluating the probable future results generated by by the marketing of this product. The projection of an income statement is therefore carried out for each future year The projection of a profit and loss account is therefore carried out for each year to come until the end of the commitments made by the insurer or over a sufficiently long period at the end of which the remaining commitments are negligible. The projection period used in our case is 50 years.
This is a long and time-consuming process, hence our desire to set up a reliable and rapid alternative, capable of serving as a device for steering sensitivities and aiding decision-making, using Machine Learning.

Machine Learning is an artificial intelligence technology that allows algorithms to learn without being explicitly to learn without having been explicitly programmed for this purpose. Machine Learning algorithms, more precisely supervised learning algorithms, will be used in a predictive approach in order to estimate the VIF. We will provide the algorithm with the data and hypotheses of the Active/Passive model and our Machine Learning model will try to reproduce the latter and restore the VIF. The image below summarizes our approach in a simple way :

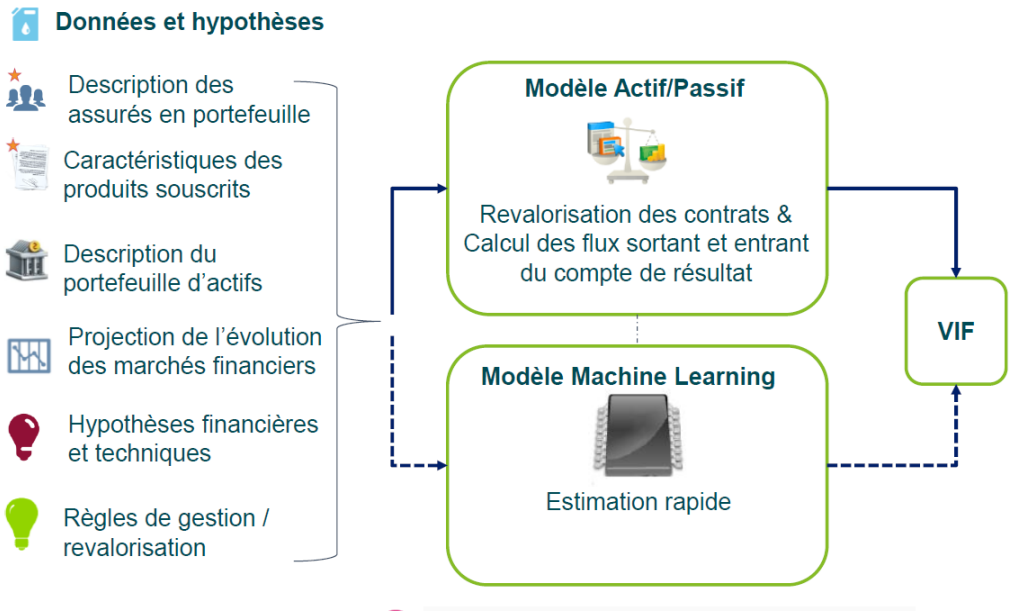

FIGURE 4 – Approche Machine Learning

The implementation of a Machine Learning model is based on 4 main steps :

— Construction of the database

— Selection of the models and scoring techniques

— Selection of the variables

— Improvement and validation of the model

We built our database based on the simulations performed by the internal modeling tool of CNP Assurances. We incorporated in our database the elements taken in-put in the model and the VIF that came out.
We have chosen to build the learning base by performing several sensitivity tests on economic, technical and financial assumptions. The objective being to constitute a reliable base allowing to capture the sensitivity and the behavior of the Asset/Liability model in different environments and configurations, we consider the observations relative to each stochastic scenario.
This database is divided into three categories : economic environment data from the GSE, contractual data (relating to the contract signed by the insured) and liabilities, and data relating to the composition of the asset portfolio.

The models chosen to address this problem are the XGBoost, the random forest, and the LASSO regression.The first two were chosen for their reputation in the field and their ability to capture complex phenomena. The LASSO regression was chosen in order to propose a simple and easily understandable

method, despite the hypothesis of the existence of a linear relationship between the explanatory variables and the variable to be explained.

We proposed a direct approach which consists in predicting directly the VIF and an approach by decomposition in sub-variables, which consists in predicting individually the three main variables which constitute the VIF and then to aggregate them. Indeed, the VIF is decomposed as follows :

$$VIF = Levies - Loads - Cost$$

Before implementing our models, we performed a variable selection. This step consists in choosing the most relevant variables for the final model and allows to reduce the complexity of the model and to improve the interpretability. To do this, we first observed the correlations between our variables as shown in the graph below :

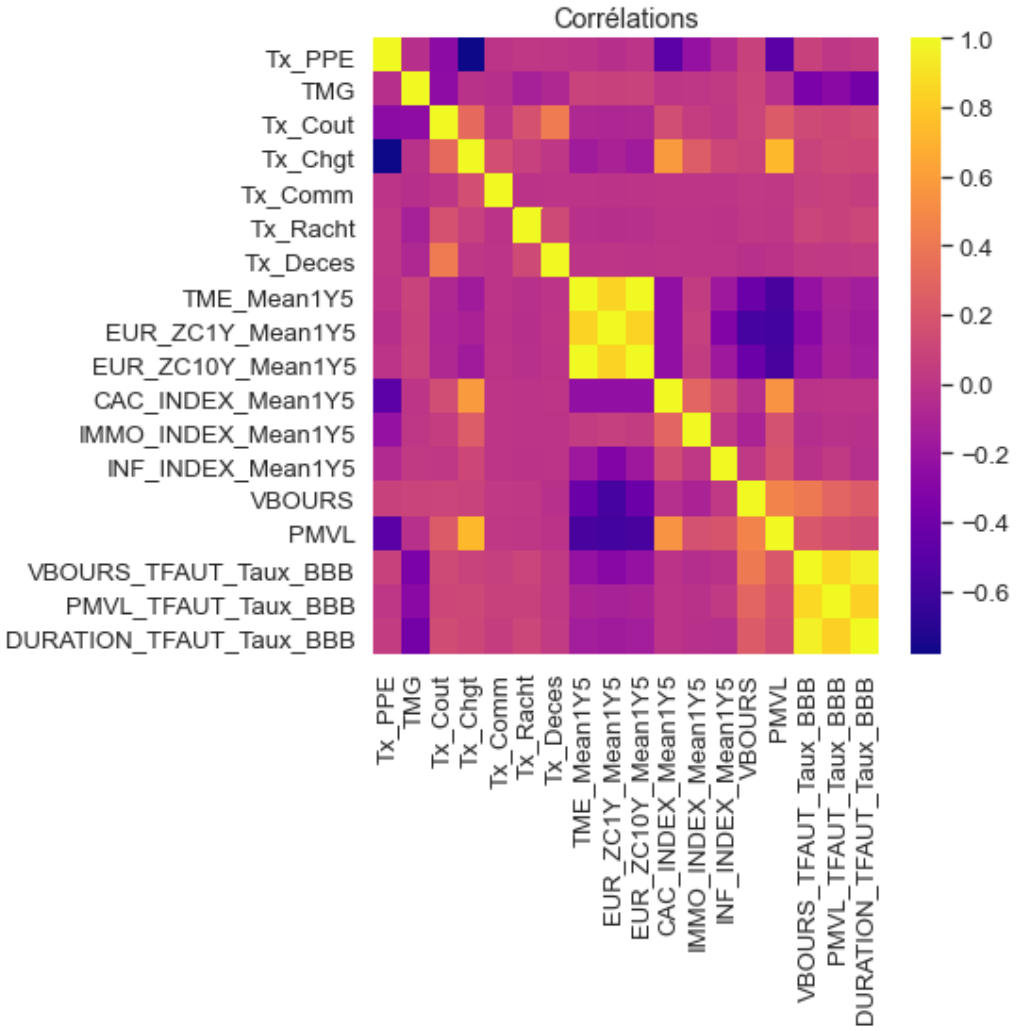

FIGURE 5 – Corrélation des variables explicatives

The observation of the correlations shows a strong correlation between the MER and zero-coupon ; between the duration, the market value (VBOURS) and the unrealized capital gains or losses (PMVL) per group of assets.
Then, the feature importance module of the XGBoost and random forest models allowed us to observe which variables were the most important. The LASSO coefficients also gave us a vision on the variables that were used in the model or not. We therefore decided to keep only the EMR and duration for the

rest of the exercise.

Once the selection of observations was done, we were able to set up our different models. We optimize the parameters of each model by using the cross-validation method which allows us to choose the most optimal a priori parameters for each model.

Below are the first results at the global scale :

Modèle	R2	RMSE (En M)	MRE	MAE (En M)
LASSO	0,75	1 850	180%	1423
RF	0,92	545	47%	238
XGB	0,99	289	15,7%	85

TABLE 7 – Résultats à une maille globale I

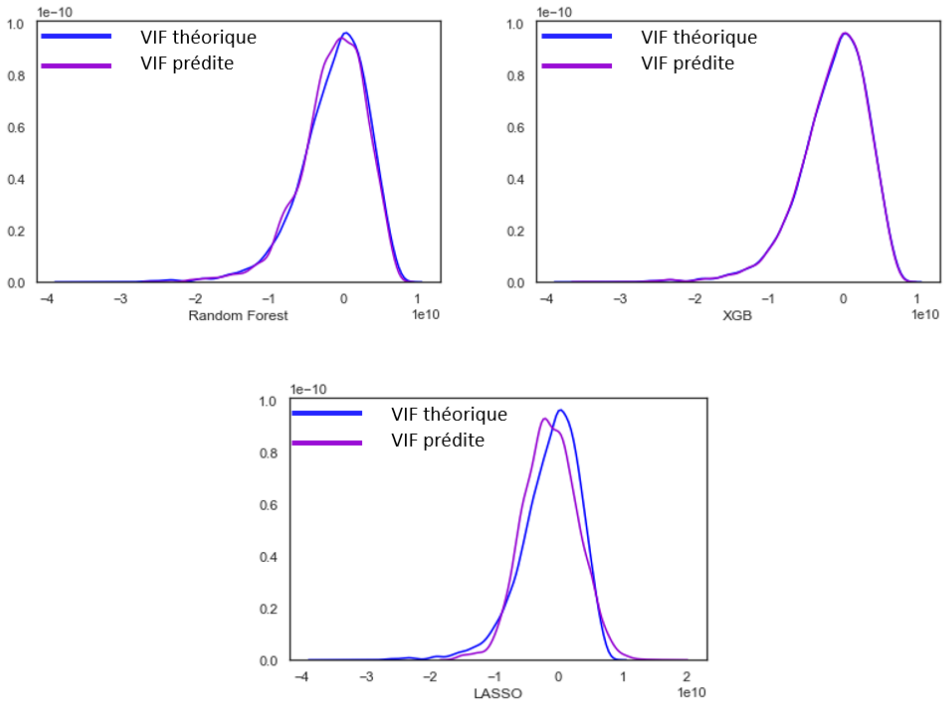

FIGURE 6 – Comparaison des densités I

These results show us that the XGBoost model is better than the others, with a density that perfectly matches the theoretical density of the VIF. It has a very high R2 of 0.99, so the model manages to correctly guess the distribution of our variable. Its average absolute relative error of 15.7% means that for each predicted observation, the model makes an error of about 15.7% more or less on the Equity. This deviation translates into a difference of €85M more or less.

We considered these results unsatisfactory and undertook to improve our results by performing a number of sensitivities, allowing us to capture the behavior of our Asset/Liability model in different configurations impacts certain key assumptions such as the level of charges, commissions, costs, etc...

Below are the new results that proved to be more conclusive :

Modèle	R2	RMSE (en M)	MRE	MAE (En M)
LASSO	0,83	1 702	180%	1 337
RF	0,99	83	9,7%	58
XGB	0,99	75	7,2%	48

TABLE 8 – Résultats à une maille globale II

The results can also be observed at a finer scale, by type of shock on the Closing :

TypeChocSCR	R2 LASSO	R2 RF	R2 XGB	RE LASSO	RE RF	RE XGB
Central	0,83	0,99	0,99	6,0%	2,5%	2,3%
ChocActionType1	0,79	0,99	0,99	-10,3%	-0,2%	0,0%
ChocActionType2	0,84	0,99	0,99	-10,4%	0,1%	0,1%
ChocBaisseRachat	0,50	0,99	0,99	10,0%	3,1%	2,9%
ChocBaisseTaux	0,79	0,99	0,99	-8,1%	2,5%	-2,7%
ChocFrais	0,81	0,99	0,99	-10,2%	-0,8%	-0,2%
ChocHausseRachat	0,87	0,99	0,99	1,7%	0,0%	-0,1%
ChocHausseTaux	0,77	0,99	0,99	-3,6%	-0,2%	-0,2%
ChocImmobilier	0,81	0,99	0,99	-13,0%	-0,6%	-0,6%
ChocLongevite	0,80	0,99	0,99	17,2%	0,0%	-0,4%
ChocMortalite	0,83	0,99	0,99	8,6%	0,3%	0,0%

TABLE 9 – Résulats à une maille Closing et par choc II

We note that each of our models has improved, in particular the Random Forest which has gone from an RMSE of 545 M to an RMSE of 83 M. The XGBoost remains the best, with an RMSE of 75 M and an MRE of 7.2%. Its relative errors on the different types of shocks are globally lower than those of the Random Forest.

We then used on our best performing model (XGBoost) a method of observation selection that allowed us to reduce the error of the model. We observed for example that the relative error on the central Closing went from 2.3% to 1.We used the sub-variable decomposition approach, which gave us the following results :

TypeChocSCR	R2 XGB	R2 ssVar	RE XGB	RE SS VAR
Central	0,99	0,99	1,6%	1,0%

TABLE 10 – Résulats à une maille Closing et par choc IV

Modèle	R2	RMSE (en M)	MRE	MAE (en M)
VIF XGB	0,99	34,8	2,6%	20,3
VIF Ss variables	0,99	28,0	1,4%	15,4
Coûts	0,99	5,7	0,2%	3,3
Commissions	0,99	8,5	0,6%	4,6
Prélèvements	0,99	13,4	0,7%	7,9

TABLE 11 – Résultats à une maille globale IV

Separating our target variable into 3 less complex variables allowed us to slightly improve our score, going from a relative error of 1.3% to 1.0% on the central VIF. We can observe that the model manages to perfectly capture the density of the theoretical variable and the R2(0.99) shows that the model understands well the hidden complexity in the computation of the VIF.
The method by sub-variable with the XGBoost model is thus the one which offers better results.

We applied the SHAP transparency algorithm to our model in order to better appreciate the results. This module allowed us to obtain additional information about the decision rules of the model and the way the VIF is estimated.

In order to apply our work, we have calculated the market SCR and the life underwriting SCR. Below are the results obtained :

SCR calculé	Valeur théorique (en M)	Valeur prédite (en M)	Ecart (en M)	Erreur relative
SCR marché	60,7	60,0	0,69	1,1%
SCR vie	4 076	4 113	-37,23	-0,9%

TABLE 12 – Résultats du calcul des SCR par module

We observe errors of 1.1% and -0.9%, which is a rather satisfactory result. The model we have set up seems to be relatively reliable in terms of results.

Our exploratory approach has convinced us that there are still several avenues of investigation to be explored in order to improve the results. Indeed, this type of modeling is not without risk of error, with biases that can come from different sources, such as the risk of expert opinion, selection, omitted data and variables.
The learning base could be enriched with new sensitivities and better methods could be found to incorporate the investment strategy and the management and revaluation rules specific to the portfolio under study.
It could be possible to test other Machine Learning models such as neural networks in order to observe their contribution.

Table des matières

Introduction

Dans un environnement multinorme en pleine évolution réglementaire, les assureurs doivent prendre en considération la multiplication des normes reflétant la réalité de manière différente et permettant une meilleure compréhension de la création de valeur. Solvabilité 2 pour la protection des assurés, IFRS 17 pour mieux refléter la fidélité des comptes et MCEV d'un point de vue actionnaire. Les assureurs doivent également faire face à des exigences de reporting de plus en plus élevées (Marchés financiers, Régulateurs, Assurés, Actionnaires et Concurrents).

Pour répondre à toutes ces exigences, les assureurs vie se doivent de calculer de nombreux indicateurs de rentabilité ajustée au risque de plus en plus sophistiqués. Dans cette étude nous nous intéressons à la valeur d'un portefeuille ou VIF (Value in Force) au sens de Solvabilité 2 d'un assureur vie proposant un produit d'épargne en euros.

Le processus de production de cette métrique demeure un exercice compliqué et chronophage. En effet, les modèles actuariels Actif/Passif utilisés par les assureurs vie reposent sur l'exploration d'un nombre élevé de scénarios via les approches stochastiques impliquant, la mise en place de nombreuses simulations. Ces modèles nécessitent ainsi en entrée un ensemble d'informations riches et variées tant sur l'environnement économique, tant sur la composition d'actifs et le passif mais aussi sur les règles de management et principes de décision de l'assureur, rendant les calculs complexes.

L'objet de cette étude portera sur le développement d'un outil de pilotage et de sensibilités, qui aura pour but de répliquer dans une certaine mesure le comportement d'un modèle actuariel Actif/Passif. Ce dispositif n'a pas pour objectif de remplacer les calculs actuels mais de permettre une estimation rapide des métriques considérées et in-fine une anticipation des résultats lors des arrêtes trimestriels.

La démarche de cette étude consiste à apprendre une variable à expliquer à partir de variables explicatives sans être explicitement programmées. Des modèles de Machine Learning ont été développés pour apprendre la valeur de portefeuille en fonction de plusieurs configurations par exemple l'environnement économique et d'un ensemble d'hypothèses techniques et financières.

Nous avons retenu trois modèles pour cette étude : les méthodes d'agrégation des arbres de régressions notamment le Random Forests et le XGBoost et la régression LASSO. Les deux premiers ont été choisis pour leurs réputation dans le domaine et leur aptitude à capter des phénomènes complexes. La régression LASSO a été choisie afin de proposer un moyen simple et facilement compréhensible malgré l'hypothèse de l'existence d'une relation linéaire entre les variables explicatives et la variable

à expliquer.

Dans ce mémoire, nous chercherons donc à fournir une solution stable, fiable et rapide pour prédire la VIF à l'aide des modèles de Machine Learning. Nos travaux se présenteront suivant cinq étapes à savoir la définition du contexte, la présentation des algorithmes de Machine Learning, la description du modèle épargne, la présentation de la base d'apprentissage et enfin l'analyse des résultats.

Chapitre 1

Contexte

1 Un environnement multinorme en pleine évolution

Les normes en assurances évoluent constamment pour s'adapter à l'évolution des situations économiques et sociales. En effet, les compagnies d'assurance sont soumises à diverses réglementations qui leur imposent un cadre spécifique en matière de souscription de contrats, de gestion des risques, de traitement des sinistres et de communication avec les clients.

On observe par exemple la nouvelle norme IFRS 17 qui est entrée en vigueur en janvier 2023 pour mieux refléter la fidélité des comptes, ou le règlement Taxonomie mis en application dans la même période pour garantir la durabilité des actifs de l'assureur.

Ces normes sont souvent dictées par des organismes de réglementation tels que l'EIOPA (Autorité européenne des assurances et des pensions professionnelles) ou l'ACPR (Autorité de contrôle prudentiel et de résolution). Leur but est de protéger les clients en garantissant la solvabilité des assureurs, la transparence des produits proposés et la qualité de services fournis.

Les normes en assurances sont également influencées par les tendances du marché et les évolutions technologiques. On a assisté à des périodes de taux bas, de taux hauts, d'inflation. Les assureurs doivent donc constamment s'adapter aux multiples changements économiques et réglementaires, d'où la montée en puissance des méthodologies et outils pour anticiper la rentabilité et le risque en assurance.

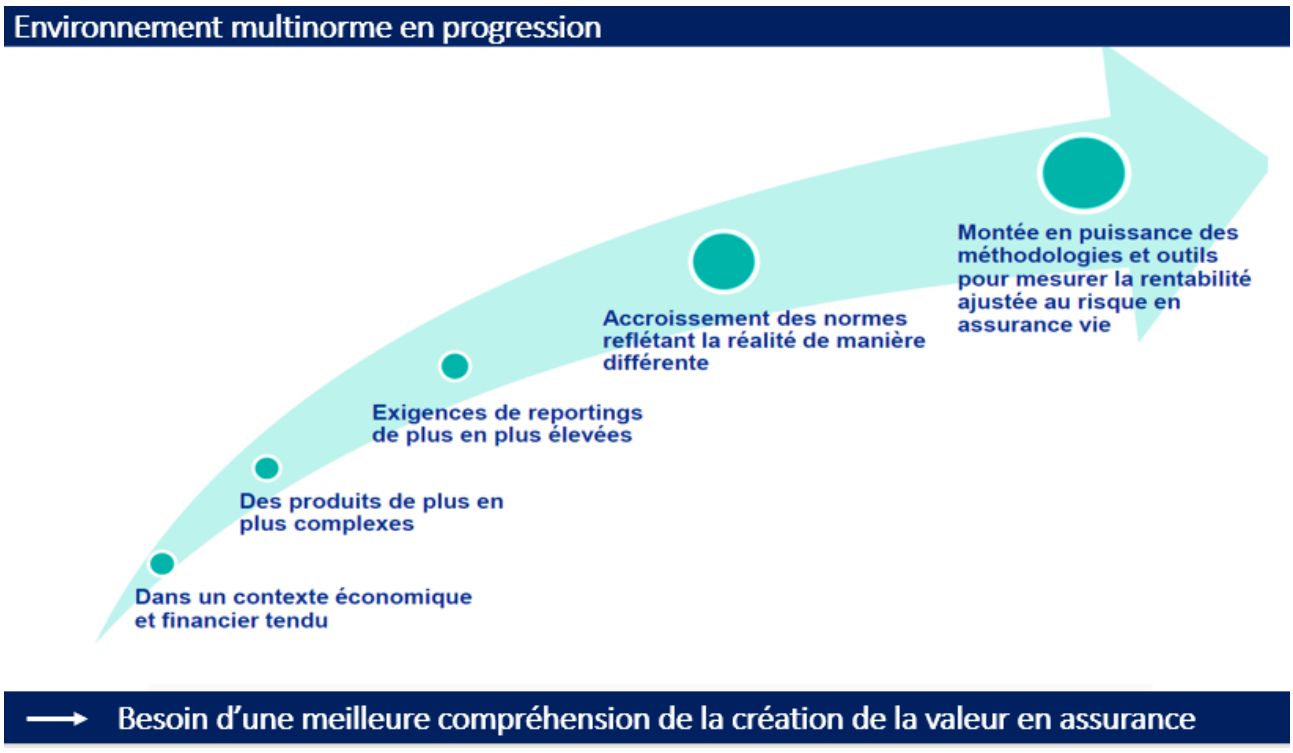

FIGURE 1.1 – Environnement multinorme en évolution

En somme, le contexte de progrssion en assurances est marqué par une double dynamique : d'une part, une adaptation constante aux exigences des régulateurs et des marchés ; d'autre part, une prise en compte croissante des enjeux sociaux et environnementaux dans la gestion des risques.

Parmi les nombreux moyens à notre disposition pour une estimation rapide de la rentabilité de l'assureur, on peut nommer le Machine Learning. En effet, ces techniques ont récemment connu une grande montée en puissance et ont fait leurs preuves dans plusieurs domaines, y compris celui de l'assurance.

2 Solvabilité II

L'assurance représente un secteur très particulier de l'économie en raison de son cycle de production inversé. En effet, contrairement à la plupart des services vendus, en contrat d'assurance, l'assureur fixe le prix de vente d'un produit (prime) avant de savoir combien cela lui coûtera (coût du sinistre). Il ne connaît donc pas à l'avance son prix d'achat, ce qui complexifie le calcul du bénéfice. De ce fait, il est important de s'assurer que l'entreprise sera en mesure de gérer tout le capital constitué par les primes et de dédommager tout client en cas de sinistre. C'est dans ce but que l'EIOPA, superviseur assurantiel européen a mis en place les normes de Solvabilité 2.

2.1 Présentation générale

La solvabilité 2 est une réforme réglementaire européenne du monde de l'assurance mise en place entre 1990 et 2010, et entrée en vigueur en 2016. Cette directive a pour but d'harmoniser la réglementation dans l'Union, d'accroître la transparence de la communication financière des assureurs et de garantir leur aptitude à honorer les engagements pris envers les personnes assurées. Sur ce dernier point, l'objectif principal est l'adéquation entre les risques liés à l'activité d'assurance et le capital alloué pour couvrir ces mêmes risques.

Chaque assureur se doit de disposer de fonds (actions, obligations, immobilier) suffisants pour faire face à des événements imprévus pouvant impacter le respect de leurs engagements. Les primes confiées par les assurés sont investis dans ces fonds et leur valeur doit être importante et le risque mesuré. Afin de mener à bien ces missions, la norme de Solvabilité 2 se répartit en 3 piliers fondamentaux comme le présente ce schéma :

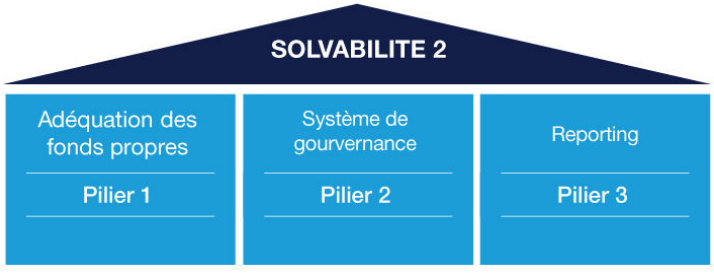

FIGURE 1.2 – Tableau expérimental des piliers de solvabilité 2

A) **Pilier 1 : Exigence de capital**
Il permet aux assureurs de retranscrire leur exposition au risque et leur gestion de capital à travers les points suivants :
- Le calcul des provisions techniques en Best Estimate (BE)
- L'élaboration d'un bilan prudentiel en valeur de marché

— Des principes en matière d'allocation et d'exigibilité des actifs

— Deux exigences de capital : le MCR (capital minimum requis) et le SCR (capital de solvabilité requis)

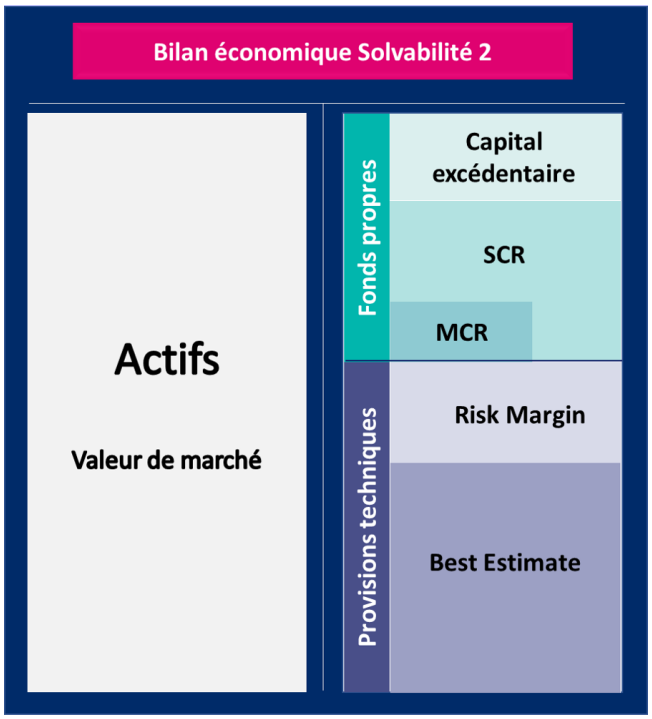

FIGURE 1.3 – Bilan économique sous Solvabilité 2

B) **Pilier 2 : Exigences qualitatives**

Ce pilier implique la mise en place d'un dispositif interne de maîtrise de tous les risques (financiers, techniques, opérationnels) auxquels peut être confronté un assureur ; pour qu'à tout moment il puisse avoir une vision précise de l'état de sa solvabilité. Les régulateurs nationaux contrôleront ces systèmes de maîtrise des risques, qui reposent sur les principes suivants :

— Mise en place de fonctions clés (actuariat, audit interne, conformité et gestion des risques)

— Gouvernance saine, prudente et effective

— Politique de qualité des données

— Mise en place d'un dispositif interne de maîtrise des risques ORSA (Own Risk and Solvency Assessment)

C) **Pilier 3 : Communication financière**

Le pilier III traite de la publication des informations sur lesquelles les deux précédents piliers sont basés et qui permettront au public (actionnaires et analystes) et aux autorités de contrôle de juger si l'analyse effectuée est fidèle à la réalité. Les assureurs et réassureurs auront donc à fournir les informations clés nécessaires à la détermination de leur exigence de capital. Ces informations devront, en particulier, couvrir les éléments suivants :

— La performance financière

— Les profils de risques, données et hypothèses sur lesquelles ils sont basés

— Les mesures d'incertitudes, incluant mesure d'adéquation des estimations antérieures et la sensibilité des résultats à la volatilité du marché...

Dans le cadre de notre étude, nous nous pencherons uniquement sur le pilier 1, celui des exigences quantitatives. Ce mémoire n'étant pas un mémoire consacré à Solvabilité 2, nous ne présenterons par la suite que les notions principales.

2.2 Les provisions techniques sous S2

La norme S2 prévoit une répartition des provisions techniques d'un assureur en deux parts qui sont le BE (Best Estimate) et la RM (Risk Margin).

- **Le Best Estimate**

Le Best Estimate ou meilleure estimation est défini par l'EIOPA comme « la meilleure estimation qui correspond à la moyenne pondérée par leur probabilité des flux de trésorerie futurs, compte tenu de la valeur temporelle de l'argent (valeur actuelle attendue des flux de trésorerie futurs), estimée sur la base de la courbe des taux sans risque pertinents .»

Le BE correspond à l'évaluation de la valeur de marché des engagements de l'assureur. Mathématiquement, il s'agit de l'espérance des flux financiers sortant diminués des flux financiers entrant. Le calcul du BE se fait à travers la formule suivante :

$$BE = \mathbb{E}\left[\sum_{t\geq 1} \frac{F_t}{(1+r(t))^t}\right]$$

Avec,

— $r(t)$ le taux d'actualisation à l'année t

— F_t la projection des flux à l'année t

- **La Marge pour risque**

La marge pour risque ou Risk Margin (RM) correspond à la valeur qu'il faut ajouter au BE pour garantir que l'ensemble des provisions techniques est équivalent au montant que les entreprises d'assurance et de réassurance demanderait pour reprendre et honorer les engagements. Elle représente donc le coût de portage du capital de solvabilité tout au long de la vie des contrats d'assurance en question. Elle a pour principaux objectifs de :

— Garantir la possibilité de transférer le portefeuille à un tiers avec un niveau de confiance suffisant

— Donner une marge de prudence suffisante pour contenir l'incertitude liée au calcul du BE

Le calcul de la RM se fait à partir de la méthode du coût du capital :

$$RM = COC \times \sum_{t\geq 0} \frac{EOF(t)}{(1+r_{t+1})^{t+1}} = COC \times \sum_{t\geq 0} \frac{SCR(t)}{(1+r_{t+1})^{t+1}}$$

Avec,

— RM la marge pour risque

— COC le taux de coût du capital. Le coût du capital correspond au taux annuel appliqué au capital requis à chaque période. Le coût du capital à utiliser s'élève à 6%.

— $EOF(t)$ les fonds propres éligibles de la compagnie après t années pour faire face à ses obligations réglementaires

— $SCR(t)$ le capital requis de la compagnie après t années.

— r_t le taux d'intérêt sans risque de maturité t

2.3 Les exigences de capital sous Solvabilité 2

La réglementation S2 prévoit 2 exigences de capital dans le passif d'un assureur. Il s'agit du SCR et du MCR.

2.3.1 Le SCR

Le SCR (Solvency Captial Requirement) correspond au niveau de capital nécessaire pour ne pas être en ruine à l'horizon d'un an avec une probabilité de 99.5%. Le SCR assure donc la couverture à 99.5% de tous les potentiels risques auxquels l'assureur pourrait être exposé dans un intervalle de un an. L'EIOPA offre aux assureurs deux méthodes pour le calcul du SCR :

— Une formule standard communiquée par l'EIOPA

— Un modèle interne développé par l'assureur et validé par l'ACPR

Dans le cadre de notre mémoire, nous ne présenterons que la formule standard proposée par l'EIOPA. La formule standard de calcul du SCR repose sur une approche modulaire, les modules correspondant aux familles de risques auxquels fait face l'assureur. Les 6 principaux modules sont les suivants :

— Le risque de marché

— Le risque de souscription en santé

— Le risque de défaut

— Le risque de vie

— Le risque de souscription en non-vie

— Le risque intangible

Ensuite, chaque module de risques est décomposé en sous modules afin d'avoir une vision plus fine et plus précise des risques encourus par l'assureur. Le tableau ci-dessous fait un récapitulatif des risques à couvrir et qui sont pris en compte par le SCR.

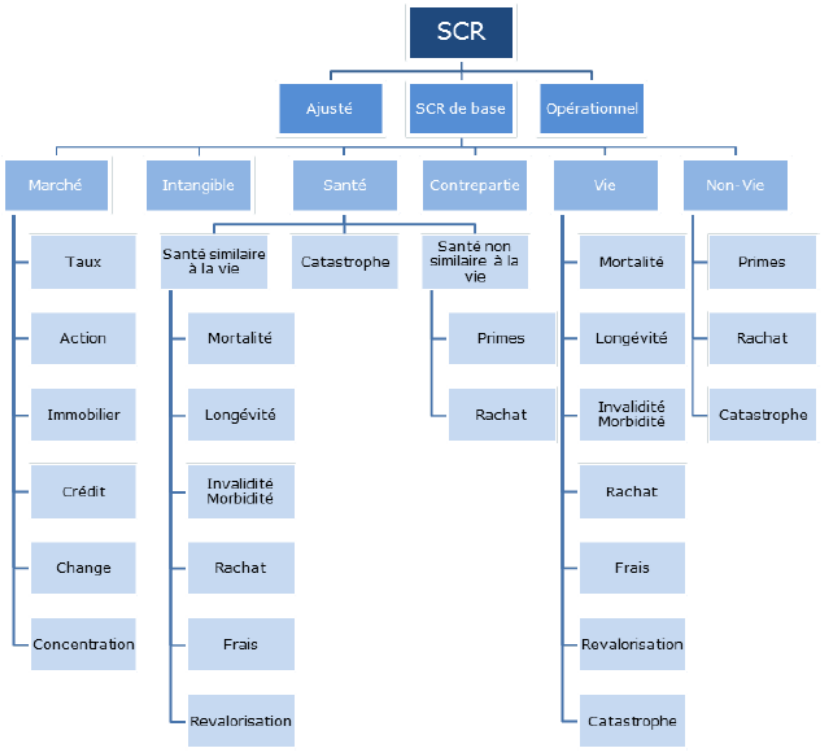

FIGURE 1.4 – Architecture modulaire du calcul du SCR en formule standard

Pour chaque module, le calcul du SCR consiste à évaluer la part des fonds propres nécessaires pour faire face à des situations des crises. Ces situations sont générées en effectuant des chocs prédéfinis par l'EIOPA, sur chaque sous module. Il peut s'agir d'une augmentation du taux de mortalité, ou encore d'une baisse des taux d'intérêts.

Ensuite, le SCR par module est calculé en faisant une agrégation des montants de fonds propres supplémentaires nécessaires via une matrice de corrélation. La formule de calcul du SCR par module est la suivante :

$$SCR_m = \sqrt{\sum_{i,j} \rho_{(i,j)} C_i Cj}$$

Avec,

- SCR_m le capital associé au module m
- C_i et C_j les capitaux associés aux risques respectifs i et j
- $\rho_{(i,j)}$ la matrice de corrélation des sous modules considérés

Puis, le BSCR (Basic SCR) est calculé en agrégeant les SCR par modules et en rajoutant le SCR dncorporel :

$$BSCR = SCR_{incorporel} + \sqrt{\sum_{i,j} Corr_{i,j} \times SCR_i \times SCR_j}$$

Enfin, le SCR total est obtenu en rajoutant au BSCR les ajustements et le SCR opérationnel :

$$SCR = BSCR + Adj + SCR_{op}$$

Avec,

— Adj l'ajustement au titre de la capacité d'absorption des pertes des provisions techniques et des impôts différés.

— SCR_{op} l'exigence de capital pour risque opérationnel

Cette formule proposée par l'EIOPA a l'avantage d'être simple à mettre en place et facilement généralisable pour tous les assureurs, bien qu'elle fasse beaucoup d'hypothèses qui ne reflètent pas forcément le risque individuel de l'assureur. Les matrices de corrélation par module, la matrice de corrélation intra-modulaire et les chocs prédéfinis par l'EIOPA sont tous disponibles dans le QIS5 publié en 2010.

- **Calcul d'un SCR sous modulaire**

Comme mentionné plus haut, avant de pouvoir agréger les modules de risque avec la matrice de corrélation spécifique et d'obtenir le BSCR, il est nécessaire de calculer les SCR sous-modulaires. Ces derniers seront eux-mêmes agrégés avec une matrice de corrélation intra-modulaire afin de donner le montant du SCR de chaque module.

Ainsi, pour chaque sous-module qui constitue un facteur de risque, nous calculons l'exigence de capital correspondante. Cette exigence de capital requis est calculée en appliquant un choc instantané (spécifié par le régulateur) sur le facteur de risque. Le calibrage des chocs retenu doit correspondre à la VaR (Value at Risk) à 0,5 % sur un an, ce qui équivaut à un scénario extrême se produisant tous les 200 ans.

Le montant du SCR pour un facteur de risque donné correspond à l'impact de ce choc sur la NAV (Net Asset Value) par rapport au scénario central du bilan, c'est-à-dire :

$$SCR_{sous-module} = NAV_{central} - NAV_{choc}$$

Avec,

— $NAV_{central}$ la NAV observée en scénario central

— NAV_{choc} la NAV observée en scénario choqué

Rappelons que la NAV (ou fonds propres) est la différence, en valeur économique, entre l'actif et le passif. De plus, le besoin en capital n'existe que si le choc détériore la situation de l'assureur, donc si $\Delta NAV > 0$, sinon nous considérons que le SCR est nul.
L'image ci-dessous explique de façon simple le procédé.

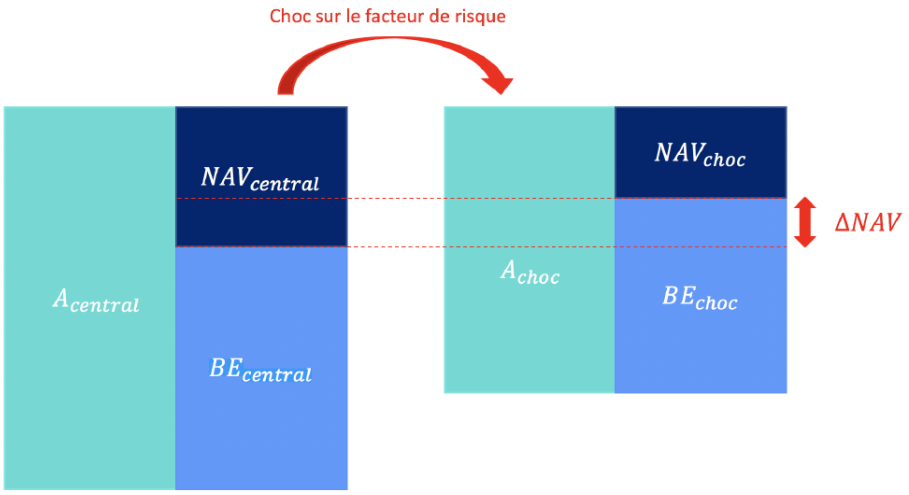

FIGURE 1.5 – Impact d'un choc sur les fonds propres économiques

Une fois les montants de tous les SCR sous-modulaires obtenus (après application des chocs imposés par le régulateur), ils seront agrégés au moyen de la matrice de corrélation intra-modulaire pour le risque donné.
Les résultats de ces agrégations produisent alors les six SCR modulaires, qui seront eux-mêmes agrégés afin d'obtenir le BSCR.

2.3.2 Le MCR

Le MCR (Minimum Capital Requirement) est le montant de fonds propres de base éligibles en deçà duquel l'entreprise d'assurance ou de réassurance court un risque inacceptable en poursuivant son activité. C'est donc la quantité de fond propre minimale exigée pour exercer légalement une activité d'assurance. Le MCR a une valeur comprise entre 25% et 45% du SCR. En dessous de ce niveau, l'intervention de l'autorité de contrôle sera automatique.

Le graphe ci-dessous résume le concept des exigences de capital sous Solvabilité 2.

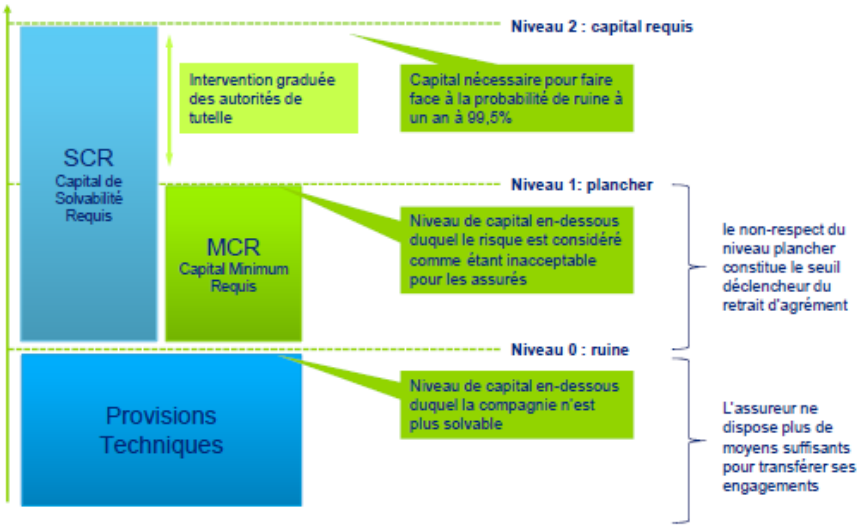

FIGURE 1.6 – Exigences de capital sous Solvabilité 2

2.4 Le ratio de solvabilité

Afin de mesurer la solvabilité d'un assureur, c'est-à-dire sa capacité à faire face à ses engagements, il est possible de mettre en place des indicateurs financiers permettant de juger la capacité de l'assureur à tenir ses promesses pendant la durée du contrat. L'un des indicateurs les plus fréquents est le ratio de solvabilité.

Le ratio de Solvabilité est défini comme le rapport des Fonds Propres (FP) de l'assureur sur son SCR :

$$RatiodeSolvabilité = \frac{FP}{SCR}$$

Un ratio supérieur ou égal à 1 signifie que l'assureur est en mesure de remplir les engagements qu'il a pris envers l'assuré. Plus le ratio de solvabilité est élevé, plus l'assureur a la possibilité d'accepter de nouveaux risques. Ce ratio s'exprime en pourcentage et les assureurs cherchent donc à obtenir un ratio supérieur à 100%.

Cet indicateur ne permet de transcrire que la solvabilité de l'assureur, c'est-à-dire sa capacité à respecter les engagements pris envers les assurés. Hors, l'assureur a également pour but de dégager des profits à partir de ses contrats d'assurance, d'être rentable. La performance d'une compagnie d'assurance repose donc sur un équilibre entre solvabilité et rentabilité.

De même que pour mesurer la solvabilité d'un assureur, il existe des indicateurs pour mesurer sa rentabilité. Parmi ceux-ci, l'un des plus courant et des plus utilisés est la VIF (Value In Force).
L'estimation de cet indicateur fera l'objet de notre mémoire.

3 La VIF comme mesure de rentabilité

3.1 Définition

Bien que la VIF ait été introduite dans le cadre du calcul de la MCEV [1], cet indicateur peut être adapté à l'étude de rentabilité en vision Solvabilité 2. La VIF correspond à la valeur actuelle probable des résultats futurs, nets d'impôts, générés par le portefeuille de contrats en cours. Elle représente la quantité de richesse générée dans le futur à partir des contrats en cours au sein du portefeuille et revenant à l'assureur.

Le calcul de la rentabilité d'un produit se fait en évaluant les résultats futurs probables générés par la commercialisation de ce produit. La projection d'un compte de résultat est donc réalisée pour chaque année à venir jusqu'à la fin des engagements pris par l'assureur ou alors sur une période suffisamment longue à l'issu de laquelle les engagements restant son négligeables. La période de projection utilisée dans notre cas est de 50 ans.

1. MCEV : La Market-Consistent Embedded Value représente une mesure de la valeur économique des activités d'assurance de personnes et des activités liées, sur la base d'une évaluation en juste valeur des actifs et des passifs. Cette métrique est établie selon les «European Insurance CFO Forum Market Consistent Embedded Value Principles »

La détermination de la VIF repose donc sur la projection des profits futurs à partir du portefeuille d'affaires actuellement en cours, le portefeuille étant déroulé jusqu'à extinction des engagements de l'assureur (on parle de run-off) en tenant compte d'hypothèses de prestations (rachats et décès).

L'Equity ou VIF est calculée de la façon suivante :

$$Equity = PNA - \sum_{n=debut\ de\ projection}^{fin\ de\ projection} VA\ frais(n)$$

Avec,

— $VA\ frais(n)$ le montant actualisé des frais payés sur la période n (frais d'entrée, de sortie, de transfert et de gestion)

Le PNA (Produit Net d'Assurance) est calculé comme suit :

$$PNA = \sum_{n=debut\ de\ projection}^{fin\ de\ projection} (VA\ prelevements\ nets(n) - VA\ commissions(n))$$
$$+Elements\ de\ fin\ de\ projection\ Assureur$$

Avec,

— $VA\ prélèvementsnets(n)$ le montant actualisé au taux sans risque des prélèvements nets sur l'année n, incluant :

- Les chargements (sur encours, sur production financière, sur primes)
- Les CFP (Coûts Fonds Propres)
- Les pénalités de rachat
- Les variations de réserve de capitalisation

— $VA\ commissions(n)$ le montant actualisé de commissions payées sur la période n (commissions sur encours, sur production financière, sur primes) augmenté du montant actualisé des commissions sur arbitrage payées sur la période n

— $Elements\ de\ fin\ de\ projection\ Assureur$ le montant actualisé des éléments de fin de projection qui reviennent à l'assureur

La VIF est une mesure intéressante car elle capture la rentabilité réelle de l'activité, c'est-à-dire la marge réelle diminuée des coûts nécessaires à l'exercice de l'activité d'assurance. Une VIF positive signifie que le contrat crée de la valeur, une richesse aux actionnaires.

On note que la VIF au sens S2 peut être définie comme la Valeur Actualisée des Profits futurs (PVFP) diminuée de la TVOG (Valeur temps des option et Granties).

$$VIF\ S2 = PVFP - TVOG$$

- **La PVFP**

La PVFP correspond à la valeur actualisée des profits futurs nets d'impôt générés par les contrats en portefeuille à la date d'évaluation et sous une hypothèse centrale d'évolution des marchés financiers alignée avec la courbe des taux de référence sur la base d'une méthodologie cohérente avec le marché.

La PVFP intègre donc la valeur intrinsèque des options et garanties financières sur les contrats en portefeuille. Les principales options et garanties financières prises en compte sont les suivantes :

— Le TMG (Taux Minnimum Garanti)

— Les garanties planchers des contrats en unités de compte

— Les options de participation aux bénéfices

— Les options de rachat

- **La TVOG**

La TVOG (Time Value of Options and Guarantees) représente le coût des options du produit modélisé. Elle peut être vue comme la différence entre le prix de l'option et sa valeur intrinsèque. En effet, la TVOG est générée par l'asymétrie de partage du sort entre actionnaires et assurés selon les diverses évolutions des marchés financiers. En absence d'options, la TVOG est nulle.

La mise en œuvre de calculs stochastiques permet, sur base de simulations multiples, de balayer le champ des possibles en termes d'évolution des marchés financiers et donc de capter le coût lié aux options financières détenues par les assurés.

La TVOG vient donc corriger le fait que la PVFP soit évaluée selon un scénario moyen. Elle se calcul comme la différence entre la PVFP Stochastique correspondant à la moyenne des PVFP_k calculées pour l'ensemble des scénarios k, et la PVFP déterministe calculée sur un scénario moyen.

$$TVOG = PVFP_{det} - PVFP_{sto}$$

4 L'estimation de la VIF avec des techniques de Machine Learning

A présent que nous avons une idée de ce qu'est la VIF et de son mode de calcul, il est important de faire le point sur deux choses : Pourquoi la VIF et pourquoi utiliser des méthodes de Machine Learning pour la prédiction de la VIF plutôt que les modèles en vigueur.
Cette partie aura pour objectif de répondre à ces deux questions en 3 parties. Nous commencerons par expliquer la nécessité de trouver une autre approche de calcul de VIF dans la problématique. Ensuite nous présenterons les raisons qui nous ont poussé à se tourner vers l'alternative du Machine Learning. Et pour finir, nous présenterons la démarche que nous allons adopter pour mener à bien ce projet.

4.1 Problématique

Dans cette étude, nous nous intéressons à la valeur d'un portefeuille d'un assureur vie proposant un produit d'épargne. La valeur du portefeuille de contrats à la date de clôture correspond à la valeur actualisée des résultats distribuables futurs après la prise en compte suffisante de risques et de

contraintes liés aux activités d'assurance dans un univers financier cohérent avec le marché.

La VIF joue un rôle crucial dans la prise de décision et le pilotage d'un portefeuille. Compte tenu du contexte financier et économique tendu auquel nous faisons face ainsi que des nouvelles réglementations mises en vigueur reflétant la réalité de manière différente, il serait bénéfique pour un assureur de comprendre rapidement l'impact de divers paramètres économiques, financiers et technique sur son activité afin de piloter l'activité et faciliter la prise de décisions.

A nos yeux, La VIF semble être la valeur adéquate à exploiter pour estimer la rentabilité. En effet, elle prend en compte une multitude de scénarios économiques possibles, ainsi que les risques des engagements pris par l'assureur jusqu'à leur extinction. Son calcul repose donc sur plusieurs hypothèses économiques et sur plusieurs années de projections, ce qui emmène les assureurs à effectuer plusieurs diverses simulations pour évaluer la totalité de leurs engagements.

C'est là tout l'objet de notre problématique : le temps. Les modèles en vigueur nécessitent en entrée un ensemble d'informations riches et variées tant sur l'environnement économique, sur la composition d'actifs et de passif mais aussi sur les hypothèses et les décisions de management rendant les calculs complexes. Le temps de production d'un exercice est susceptible de prendre jusqu'à 3 mois. Hors l'idéal serait de pouvoir à tout moment, estimer immédiatement les variations de la VIF en cas d'évolution des paramètres clés relatifs aux changements de l'environnement financier et économique.

Notre objectif est donc de développer un outil capable de répliquer le fonctionnement du modèle actuariel, et qui serait en mesure d'estimer la VIF à l'aide des différentes informations relatives au portefeuille étudié. Cet outil n'a pas pour vocation de remplacer le modèle existant mais de servir de support pour des études ad hoc et de sensibilités.

4.2 Le Machine Learning comme solution

L'une des réponses aux besoin d'accélération du calcul de la VIF serait l'utilisation du Machine Learning. C'est la piste que nous proposons d'étudier dans ce mémoire. En effet le Machine Learning, pourrait constituer un moyen d'estimation fiable et rapide de la valeur de notre portefeuille.

Le Machine Learning est une technologie de l'intelligence artificielle permettant à des algorithmes d'apprendre sans avoir été explicitement programmées à cet effet. Les algorithmes de Machine Learning, plus précisément d'apprentissage supervisé seront donc utilisés dans une démarche prédictive dans le but d'estimer la VIF. Nous fournirons à l'algorithmes les données et hypothèses du modèle Actif/Passif et notre modèle de Machine Learning devra essayer de reproduire ce dernier et de calculer la VIF.
Le schéma ci-dessous résume notre approche de façon simple.

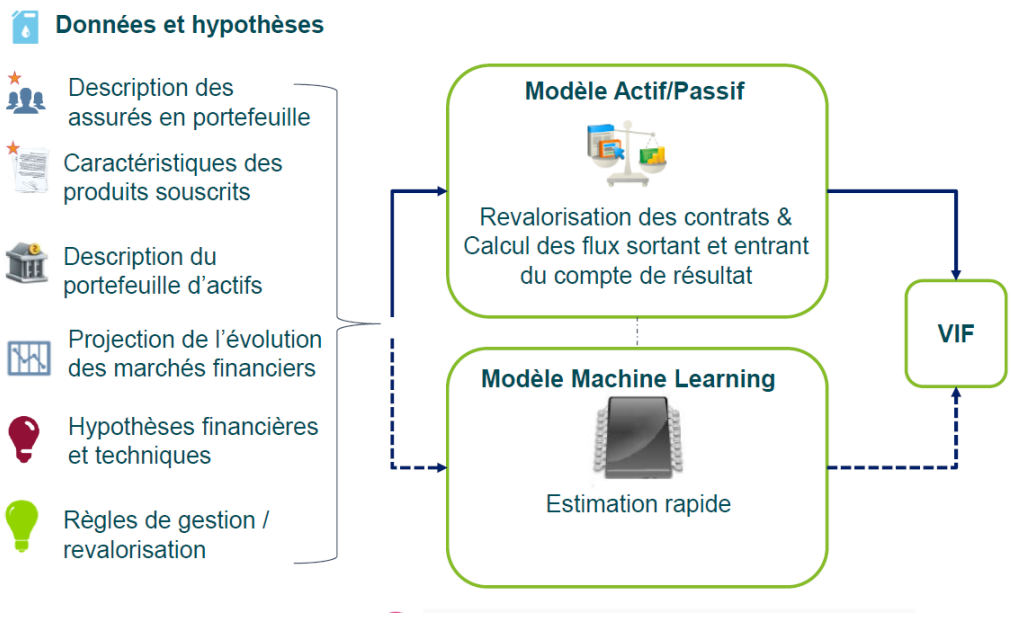

FIGURE 1.7 – Approche Machine Learning

Si nous avons décidé d'explorer des méthodes basées sur le Machine Learning, c'est pour trois principales raisons.

- **Une méthode qui a déjà fait ces preuves**

Les techniques de Machine Learning connaissent une montée en puissance depuis quelques années et sont utilisées dans de nombreux domaines variés. Ces techniques sont réputées pour leur efficacité et leur utilité dans le monde d'aujourd'hui. Il existe de nombreuses applications telles que la reconnaissance faciale sur un téléphone, le tri de spam dans une boite e-mail, la détection automatique d'un éventuel cancer sur des radios de patients, la détection de fraude bancaire,etc... Le Machine Learning permet la résolution de nombreux problème qu'on rencontre au quotidien et même dans le domaine de la banque et de l'assurance. Plusieurs problèmes statistiques ont pu être résolus grâce aux techniques de Machine Learning.

- **Un avantage sur l'apprentissage statistique**

En apprentissage statistique classique, on suppose que la variable étudiée suit une certaine loi et on essaye de déterminer les paramètres de cette loi en réduisant le bruit. C'est donc une méthode limitée par les lois à disposition. En Machine Learning, on ne pose pas d'hypothèses sur les lois, ce qui donne bien plus de possibilités d'ouvertures et de choix. Les modèles de Machine Learning sont calibrés à partir des données fournis et l'objectif est d'estimer correctement la variable cible (la VIF ici) en réduisant l'erreur de prédiction au maximum.
Nous sommes donc dans une démarche purement prédictive et non pas analytique. De plus, les avancées technologiques sur l'intelligence artificielle offrent de grandes possibilités et de bonnes probabilités de performance aux modèles de Machine Learning. L'apprentissage automatique combine à la fois les méthodes statistiques et les technologies d'intelligence artificielle.

- **La rapidité du temps de calcul**

Pour finir, l'une des raisons qui nous a emmenés à choisir le Machine Learning comme alternative est le temps. Il est vrai que la mise en place d'un modèle de Machine Learning passe par plusieurs étapes et nécessite temps, mais une fois le modèle abouti, l'exercice de prédiction s'avère être rapide. Prenons exemple sur le tri des spams. C'est quelque chose d'instantané qui se fait à la seconde où le mail arrive dans votre boîte. Il en est de même pour les modèles que nous mettrons en place. La prédiction de la VIF se fera presqu'instantanément. Cela offrira alors la possibilité de jouer sur de nombreuses sensibilités et d'avoir une réponse rapide aux questions sur la rentabilité de l'assureur.

4.3 Notre démarche

Afin de mettre en place notre modèle, nous passerons par les 4 étapes essentielles :

- **Construction de la base de données**

L'apprentissage effectué par un modèle de Machine Learning est basé sur les données que ce dernier reçoit en entrée. Celles-ci jouent donc un rôle important dans la méthodologie et impactent énormément les résultats du modèle. La construction d'une bonne base de données est donc une étape primordiale dans le cadre de notre étude. Nous la détaillerons dans une autre partie. La base de données est souvent séparée en deux parties à savoir les variables explicatives et la variable à expliquer. Les variables explicatives sont les variables qui permettront à notre modèle à apprendre tandis que la variable à expliquer est celle que l'on souhaite prédire.

- **Sélection de modèles et score**

Compte tenu de la multitude d'algorithmes de Machine Learning existant, il est impératif de faire une sélection pertinente et efficace des modèles à exécuter. Les modèles sont généralement sélectionnés en fonction d type de données et de la difficulté du problème à résoudre.
Dans notre cas, nous faisons appel dans un premier temps à deux modèles réputés pour leur bonnes performances, même face aux problème les plus complexes : Le random forest et le XGBoost. Nous utiliserons également la régression LASSO, qui est une approche simple, facilement compréhensible, et efficace lors de mise en place des modèles avec une multitudes de variables.
Ensuite pour évaluer la pertinence de ces algorithmes et afin de pouvoir les comparer les uns aux autres, il est capital de définir des fonctions de scores adaptées qui nous serviront à mesurer la performance de chaque algorithme.

- **Entraînement de l'algorithme et Sélection des variables**

Une fois que la base est construite et les modèles sont choisis, on passe à une première exécution du modèle. Pour le faire, la base de données est scindée en deux parties. Une base d'apprentissage et une base et une base de test. La base doit être découpée de façon homogène. Nous utilisons la répartition (75,25) c'est-à-dire que la base d'apprentissage constituera 75% de notre base initiale et la base de test 25%. Cette répartition est la plus utilisée mais peut varier en fonction du type de données et du modèle.

Cette séparation est faite pour évaluer la capacité du modèle à appréhender de nouvelles données. En effet le modèle apprendra sur un set de données et effectuera des prédictions sur de nouvelles données qu'il ne connaît pas encore. Il sera donc testé sur ces nouvelles données.

Une fois cette répartition appliquée, nous exécuterons nos algorithmes sur la base d'apprentissage et feront ressortir les variables qui ayant joué un grand rôle dans le calcul de la variable cible. A partir des variables choisies par le modèle comme importantes d'un regard d'expert sur le sujet, nous ferons une sélection définitive des variables à utiliser. Notre objectif est de développer un modèle transparent par design dès la phase de conception et ce en retenant uniquement les variables les plus pertinentes. Une analyse détaillée sur la sélection des variables sera effectuée dans le chapitre 5.

- **Amélioration et Validation du modèle**

La dernière étape consiste au perfectionnement du modèle. Après avoir sélectionné les variables qui offraient les résultats les plus satisfaisants dans l'étape précédente, il faut passer à l'hyper paramétrisation du modèle. Cette étape consiste à modifier les différents paramètres de modèles, afin de trouver la combinaison qui sied le mieux au modèle. Cette étape peut être effectué grâce à une méthode appelée validation croisée, que nous détaillerons plus tard. Après cette étape, on obtient une version finale du modèle, que l'on peut valider, ou réfuter, en fonction des scores qu'il présente sur de nouvelles données, de l'équilibre biais/variance des résultats, et d'autres critères de sélections.

4.4 L'approche par décomposition en sous variables

Dans le cadre de notre étude, nous avons pensé à deux approches différentes pour prédire la VIF. La première, toute simple, consiste à utiliser le modèle de Machine Learning pour prédire la VIF. Mais puisque que la VIF est une variable dont la modélisation repose sur plusieurs calculs complexes, nous avons jugé opportun de la scinder en plusieurs variables « plus simples » à estimer.

La VIF peut aussi se décomposer de la façon suivante :

$$VIF = Prélèvements - Commissions - Coûts$$

La deuxième approche consistera donc à prédire individuellement, les différentes variables qui constituent la VIF. Il s'agit notamment des variables suivantes :

— Les Coûts : Les coûts correspondent à l'ensemble des frais relatifs aux engagement présents dans le portefeuille de l'assureur. Le montant de frais pour chaque exercice est distingué selon les destinations de coûts suivantes ; frais d'acquisition, frais de gestion des sinistres et tous les autres frais de gestion. Pour chacune de ces catégories, les frais dépendront de l'évolution de différents drivers (exemple : PM, nombre de sinistres, montant de sinistre, etc.) et d'un montant unitaire pour chacun d'eux (par exemple par un coût unitaire par contrat/sinistre, un pourcentage de frais par PM/montant de sinistre, etc.). Le montant des frais par exercice est la somme des frais projetés pour chacun de ces drivers.

— Les Commissions : Les commissions représentent la rémunération des partenaires et des réseaux de distributions des contrats. On distingue plusieurs types de commissions ; les commissions sur primes, sur encours ou encore sur production financière

— Les prélèvements : On distingue plusieurs types de chargements : les chargements sur primes, sur encours ou encore sur production financière. Les chargements sur primes proviennent des conditions contractuelles et correspondent aux prélèvements effectués pour couvrir éventuellement les

frais d'acquisition et les commissions. Les chargements sur primes (souscriptions, versements libres, versements programmés) sont calculés en pourcentage des primes de la période. Les chargements sur encours ou sur production financière correspondent aux prélèvements effectués pour gérer les contrats.

On fera référence à ces variables comme des « sous variables à expliquer ».
On estime que la prédiction d'une variable telle que les Commissions est plus simple que celle de la VIF. C'est pourquoi, la deuxième démarche pourrait être meilleure que la première. Nous ferons une analyse comparative de ces deux méthodes dans la suite de ce mémoire.

Chapitre 2

L'apprentissage supervisé

1 Introduction au Machine Learning

1.1 Définition

Le Machine Learning (ML) a émergé dans la seconde moitié du XXème siècle du domaine de l'intelligence artificielle et correspond à l'élaboration d'algorithmes capables d'accumuler de la connaissance et de l'intelligence à partir d'expériences, sans être humainement guidés au cours de leur apprentissage, ni explicitement programmés pour gérer des expériences ou données spécifiques.
Arthur Samuel, considéré comme créateur du Machine Learning en 1959, le définit comme la science de donner à une machine la capacité d'apprendre, sans la programmer de façon explicite.

L'objectif du ML ou apprentissage automatique est de faire apprendre à des algorithmes des « patterns », c'est à dire une structure ou des motifs récurrents dans un ensemble de données. Il peut s'agir de chiffres, de mots, d'images, etc... Une fois que l'algorithme aura saisi la structure des données et les différents liens entre celles-ci, il sera apte à reproduire des données suivant le même schéma.

Le Machine Learning s'inscrit dans une démarche principalement prédictive, c'est-à-dire qu'il a pour objectif de déceler les relations qui lient un ensemble de variables (appelées variables explicatives) à une autre (appelée variable à expliquer). A partir de ce qu'il aura retenu en apprenant sur les premières données, il pourra prédire de façon indépendante, la variable à expliquer, à partir de nouvelles variables explicatives fournies.

Au vu des évolutions constantes et nombreuses dans le domaine de l'informatique, nous disposons de plus en plus de méthodes de traitement et de stockage de données. Cela a favorisé l'émergence et le développement du Machine Learning, qui est désormais utilisé dans de nombreux domaines tels que la Cyber sécurité, le diagnostic médical, la détection de Fraude, etc...

1.2 Les types d'apprentissage

Il existe une large variété d'algorithmes de Machine Learning mais on peut les regrouper en trois principaux groupes :

- **Les algorithmes d'apprentissage supervisé**

On parle d'apprentissage supervisé lorsque le modèle connaît déjà les réponses qu'on attend de lui, il travaille à partir de données étiquetées. Cela signifie qu'on dispose d'un ensemble de données pour apprendre (entrée) et d'une variable à prédire (sortie). Ce type d'algorithme permet d'effectuer deux types de tâches : La classification et la régression.

La classification consiste à repartir les objets dans un ensemble de classes définies. Par exemple, ce modèle sera utilisé pour savoir entre deux images soumises à l'algorithme, laquelle est celle d'un visage humain ou pas. Les sorties de classifications ne prennent qu'un nombre de valeurs finies et connues à l'avance (par exemple 0 ou 1.)

En régression en revanche, on cherche à estimer une valeur mathématique, un nombre réel. Il peut s'agir par exemple de la valeur de l'immobilier à Paris dans 15 ans ou du résultat d'un assureur comme

dans notre cas.

- **Les algorithmes d'apprentissage non supervisé**

Contrairement aux algorithmes supervisés, les données utilisées pour les algorithmes non supervisés ne sont pas étiquetées. Il n'y a pas de variable explicative à prédire et le modèle a pour objectif de créer par lui-même des groupes homogènes de données. Ensuite, il sera apte à repartir de nouvelles données dans les différents groupes qu'il aura créés. Ces techniques peuvent se montrer pertinentes dans le cas d'études sur le comportement humain. Ces méthodes sont généralement utilisées pour des problèmes de classification.

- **Les algorithmes d'apprentissage par renforcement**

Dans ce cas de figure, l'algorithme a pour objectif d'apprendre à la suite d'essais et en corrigeant ses précédentes erreurs après chaque nouvel essai et par interaction avec l'environnement.
Pour se faire, une fonction de récompense tenant compte des actions de l'algorithme est définie. L'algorithme recevra alors des bonus pour chaque bonne action et des malus pour les mauvaises. Il ajuste alors son comportement à chaque nouvel essai dans le but de maximiser la fonction de récompense.
Ce type d'apprentissage est particulièrement efficace dans les tâches où l'environnement est complexe et les règles difficiles à articuler en langage informatique. L'apprentissage par renforcement est très utile dans des domaines tels que les jeux ou la robotique.

Dans le cadre de nos travaux, nous ne nous pencherons que sur les méthodes utilisées pour nos travaux, c'est à dire les méthodes d'apprentissage supervisé.

2 Principe et notation en apprentissage supervisé

2.1 Notations et fonction de coût

L'apprentissage supervisé consiste à prédire une variable à expliquer Y à partir d'un ensemble de variables explicatives $(X^j)_{0\leq j\leq p}$ prédéfinies, X^j étant un vecteur de même dimension que Y. Si Y est continue, on parle de régression et si Y est catégorielle (ne peut prendre qu'un nombre défini de valeurs) alors il s'agit d'une classification.

On notera :

— n le nombre d'observations, donc la taille du vecteur Y

— p le nombre de variable explicatives

— X la matrice des variables explicative, de taille $n \times p$

— X^j le vecteur de taille n représentatif d'une variable explicative j

— X_i^j la valeur de la variable j pour l'observation i

— $X_i = (X_i^1, X_i^2, ..., X_i^p)$ l'ensemble des valeurs de l'observation i sur les différentes variables explicatives

L'objectif en apprentissage supervisé est de comprendre et d'apprendre les connections qu'il y'a entre chaque X_i et Y_i, afin de pouvoir de pouvoir prédire la valeur Y_m pour un nouveau X_m donné.

Pour cela, on suppose qu'il existe une fonction $f \in F$ telle que $f(X) = Y$, où F représente l'ensemble des méthodes supervisés existantes. Puisqu'on ne connaît pas la réelle nature de la relation entre X et Y et puisque les données X peuvent être bruitées ou incomplètes, il est plus judicieux de noter le problème ainsi :

$$Y = f(X) + \epsilon$$

Où, ϵ est le résidu.

On notera :

— $\hat{f}$ la fonction de lien utilisée par l'algorithme

— $\hat{Y}$ le vecteur prédit par l'algorithme, donc $\hat{f}(X) = \hat{Y}$

— ε l'erreur commise par le modèle, soit $\varepsilon = Y - \hat{Y}$. Il s'agit donc d'une valeur qu'il faudra minimiser pour obtenir un bon modèle.

La première étape pour trouver la fonction f adéquate consiste à mettre en place une fonction de coût.

La fonction de coût encore appelée fonction de perte est une fonction qui a pour objectif de mesurer l'écart entre la valeur attendue Y_i et la valeur prédite $\hat{Y}$. Plus la fonction de coût est faible, plus le modèle est performant. Elle se définit comme suit :

$$\begin{array}{rccc} L: & \mathbb{R}^n \times \mathbb{R}^n & \longrightarrow & \mathbb{R}^2 \\ & (Y, \hat{Y}) & \longmapsto & L(Y, \hat{Y}) \end{array}$$

Certains algorithmes tels que le Gradient Descent que nous verrons plus tard se servent de la fonction de coût pour optimiser leur performance, en ajustant progressivement les paramètres du modèle dans l'objectif de réduire la fonction de coût.
C'est donc un élément clé de la modélisation en Machine Learning car elle permet d'évaluer la qualité des prédictions du modèle et d'ajuster les paramètres pour améliorer ses performances.

Il existe différentes fonctions de coût, chacune étant adaptée à un type de problème spécifique. La fonction de coût la plus utilisée dans le domaine de la régression est la MSE (Mean Squarre Error), qui calcule la moyenne des écarts entre les valeurs prédites et les cibles pour chaque observation. Elle se définit comme suit :

$$L(Y, \hat{Y}) = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2$$

2.2 Erreur quadratique

L'erreur quadratique correspond à l'espérance de la somme des écarts entre la valeur observée et la valeur prédite. Elle est directement associée à la MSE. Elle se développe comme suit :

$\mathbb{E}[L(Y, \hat{Y})] = \mathbb{E}[(Y - \hat{Y})^2]$

$= \mathbb{E}[(\hat{Y} - \mathbb{E}[\hat{Y}] + \mathbb{E}[\hat{Y}] - Y)^2]$

$$= \mathbb{E}[(\hat{Y} - \mathbb{E}[\hat{Y}])^2] + \mathbb{E}[(\mathbb{E}[\hat{Y}] - Y)^2] + 2 \times \mathbb{E}[(\hat{Y} - \mathbb{E}[\hat{Y}]) \times (\mathbb{E}[\hat{Y}] - Y)]$$

$$= \mathbb{E}[(\hat{Y} - \mathbb{E}[\hat{Y}])^2] + (\mathbb{E}[\hat{Y}] - Y)^2 + 2 \times \mathbb{E}[\hat{Y} - \mathbb{E}[\hat{Y}]] \times (\mathbb{E}[\hat{Y}] - Y)$$

$$= \mathbb{E}[(\hat{Y} - \mathbb{E}[\hat{Y}])^2] + (\mathbb{E}[\hat{Y}] - Y)^2 + 0 \; car \; \mathbb{E}[\hat{Y} - \mathbb{E}[\hat{Y}]] = \mathbb{E}[\hat{Y}] - \mathbb{E}[\hat{Y}] = 0$$

$$= Var(\hat{Y}) + Biais(\hat{Y})^2$$

Où

— $Var(\hat{Y}) = \mathbb{E}[(\hat{Y} - \mathbb{E}[\hat{Y}])^2]$

— $Biais(\hat{Y})^2 = (\mathbb{E}[\hat{Y}] - Y)$

Le modèle admet donc deux fonctions qu'il faut minimiser : le biais et la variance

2.3 Le dilemme biais variance

Le biais représente l'écart observé entre la valeur prédite par le modèle et la valeur observée. Un biais élevé signifie donc que l'algorithme n'arrive pas à comprendre et à prédire les liens entre les variables d'entrée et la variable de sortie.
La présence de biais est généralement dû à l'utilisation d'un modèle trop simple ou non adéquat. Dans ce cas il est donc considéré comme des erreurs causées par des hypothèses incorrectes dans l'algorithme d'apprentissage.

Le biais peut également être introduit par les données d'entrées, dans le cas où celles-ci ne seraient pas représentatives de la population ou dans le cas où elles ne seraient simplement pas liées à la variable de sortie. Par exemple, il ne serait pas surprenant d'obtenir de mauvais résultats en essayant de prédire l'âge du fils aîné en fonction de la couleur des yeux du père. En effet les deux variables n'ont rien à voir de prime abord. La pertinence du modèle repose donc sur la mise en place d'une base de données pertinente et fiable. L'image ci-dessous illustre de façon simple la notion de biais :

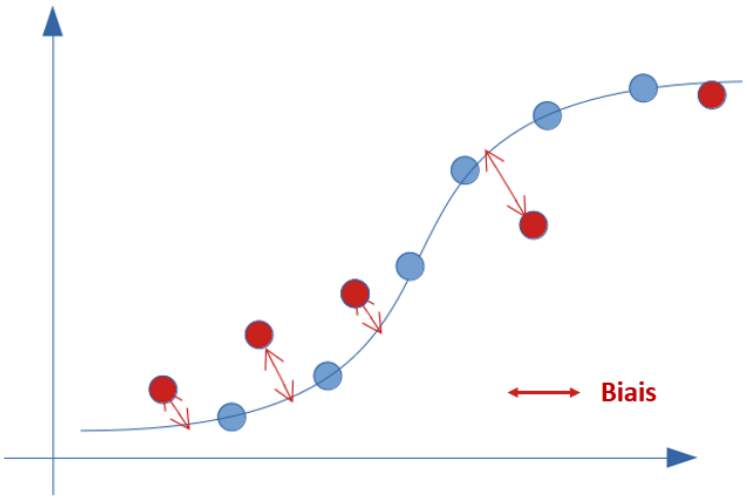

FIGURE 2.1 – Illustration du biais

En supposant que la courbe bleue représente l'ensemble des valeurs observées Y_i, et les cercles l'ensemble des valeurs prédites $\hat{Y}_i$, le biais serait alors les flèches rouges, qui montrent l'écart entre les deux.

La variance correspond à l'erreur de sensibilité des données utilisées pour le modèle d'apprentissage automatique. Elle représente la capacité du modèle à se généraliser sur de nouvelles données et est donc liée à la complexité du modèle. En effet, plus un modèle sera complexe, plus il aura tendance à vouloir apprendre le bruit aléatoire pour prédire avec exactitude les données qu'il a eu en entrée. L'apprentissage du modèle sera donc centré sur les données d'apprentissage et il sera difficile de l'utiliser sur de nouvelles données.

Une méthode courante pour diminuer la variance consiste donc à utiliser un modèle moins complexe, ce qui entraîne naturellement la dégradation du pouvoir de prédiction (les prédictions sont moins exactes) et donc la hausse du biais. Réciproquement, la réduction du biais d'un algorithme se fait au prix de l'augmentation de sa variance. On parle alors du dilemme biais-variance. Il s'agit d'une problématique centrale en apprentissage supervisé.

Un algorithme avec un biais élevé et une variance faible sera facilement généralisable à d'autres modèles malgré que son erreur élevée. Tandis qu'un modèle à biais faible et variance élevée fournira de bonnes de prédictions sur des données similaires aux données d'entrée mais les résultats seront incertains sur une nouvelle population.
L'image ci-dessus illustre le concept du dilemme biais-variance :

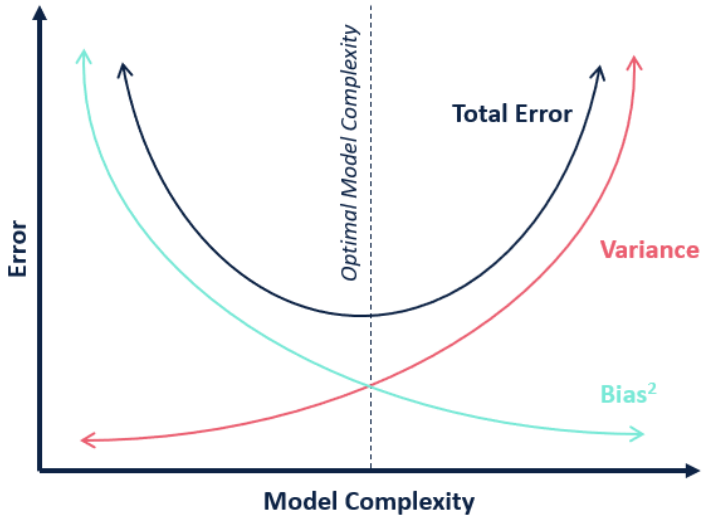

FIGURE 2.2 – Dilemme biais-variance

Comme on peut l'observer, le biais diminue tandis que la complexité et la variance du modèle augmente. De même, le biais augmente et la variance diminue lorsque le modèle baisse en complexité. Il faut donc trouver un équilibre entre un modèle trop complexe et un modèle pas assez complexe. Ceci conduit directement aux notions de sur-apprentissage et sous-apprentissage.

2.4 Surapprentissage et sous apprentissage

Le surapprentissage ou « *over-fiting* » survient lorsque le modèle apprend « trop » les données d'apprentissage. Le modèle apprend jusqu'au bruit aléatoire et ne peut donc pas être généralisable. On peut facilement déduire un cas de sur-apprentissage lorsque le modèle a un de très bon résultats de prédiction sur les données d'apprentissage mais qu'il prédit mal les données de test. C'est un modèle à variance élevé et à biais faible.

Inversement, le sous apprentissage ou « *under-fiting* » survient lorsque que le modèle n'est pas suffisamment complexe et de ce fait ne parvient pas à comprendre les relations entre les données d'entrées et les données de sortie. Il s'agit d'un modèle à biais élevé .

Le surapprentissage et le sous apprentissage sont les principales causes de mauvaises performances pour un algorithme supervisé. L'image ci-dessous illustre bien la notion de sous et sur apprentissage :

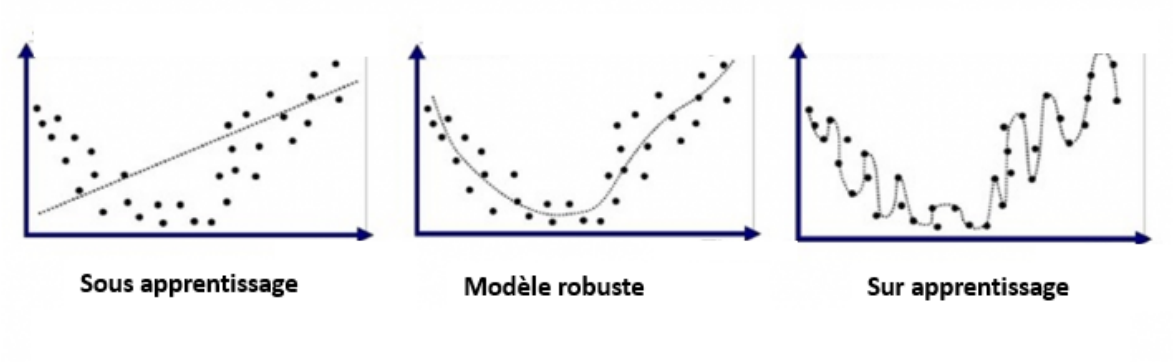

FIGURE 2.3 – Sous apprentissage et surapprentissage

3 Indicateurs de performance

Etant donné que nous effectuerons des tests sur plusieurs algorithmes de régression, il nous faut définir des mesures d'erreur applicable à chaque modèle, afin de pouvoir les comparer les uns aux autres. Ces mesures d'erreur nous serviront d'indicateur pour la sélection du modèle final et nous permettront également d'avoir une vision quant à la performance globale et la qualité d'ajustement de chaque modèle.
Les mesures d'erreur sélectionnées pour notre étude sont les suivantes :

- **R2 Score ou Coefficient de détermination de Pearson**

Il s'agit d'un indicateur bien connu en statistiques pour juger de la qualité d'un modèle de régression. On peut voir le R2 comme l'erreur du modèle divisée par l'erreur d'un modèle basique qui prédit tout le temps la moyenne de la variable à prédire.
Cet indicateur est compris entre 0 et 1, et est croissant avec la performance du modèle.
Il se calcule comme suit :

$$R2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Avec,

— $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, la moyenne des Y_i

- **MSE et RMSE**

Comme nous l'avons mentionné plus haut, la MSE est une mesure couramment utilisée en ML car elle permet d'apprécier les performances du modèle en mesurant la moyenne des écarts entre valeurs observées et valeurs prédites. De surcroît, la MSE est fonction du biais et de la variance. Ce qui signifie qu'elle nous donne à la fois la capacité du modèle à prédire la variable cible (minimiser le biais) et à se généraliser à de nouvelles données (minimiser la variance).

Cependant, la MSE a un inconvénient majeur : elle peut être affectée par les valeurs aberrantes (outliers), car elle pénalise fortement les grands écarts. Pour remédier à cela, nous utiliserons dans nos travaux la RMSE (Root Mean Squared Error). Comme son nom l'indique, la RMSE est simplement la racine carrée de la MSE :

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y - \hat{Y})^2}$$

La RMSE permet d'obtenir une mesure de l'écart moyen entre les prédictions et les véritables valeurs, qui est plus facilement interprétable en termes d'échelle la variable d'origine, tout en ayant l'avantage de limiter l'impact des valeurs aberrantes.

- **MRE**

La MRE (Mean Relative Error) ou erreur relative moyenne est un indicateur simple et efficace en Machine Learning pour évaluer la performance d'un modèle. Elle permet de mesurer l'écart en pourcentage entre la prédiction et la réalité. Cette mesure est particulièrement utile dans les domaines où les valeurs sont très variables, car elle permet de comparer les erreurs pour toutes les valeurs, quel que soit leur ordre de grandeur.

Cependant, il est important de prendre en compte les limites de l'erreur relative. Par exemple, si la valeur réelle est très proche de zéro, l'erreur relative peut devenir très grande et ne pas refléter la performance réelle du modèle.
La MRE est calculée comme suit :

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i - \hat{Y}_i|}{|Y_i|}$$

- **MAE**

La Mean Absolute Error (MAE) mesure l'erreur absolue moyenne entre les valeurs prédites et les valeurs réelles dans un ensemble de données. Il s'agit de la moyenne des biais en valeur absolue pour chaque observation prédite par le modèle
Elle se calcule comme suit :

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

Contrairement à la (RMSE), la MAE ne prend pas en compte les erreurs quadratiques. Cela signifie que la MAE est une mesure plus robuste aux erreurs extrêmes. Elle est également plus facile à interpréter que la RMSE, car elle représente directement l'erreur moyenne en termes de l'unité de mesure du problème.

Elle est facile à interpréter dans le contexte du problème. La formule de calcul est simple et directe, ce qui en fait un outil pratique pour mesurer l'efficacité des algorithmes de prédiction.

4 Présentation des modèles sélectionnés

Nous allons maintenant présenter les trois méthodes retenues pour notre démarche. Il s'agit de la régression LASSO, du XGBoost et du random forest ou forêts aléatoires. La première méthode est une technique assez basique, et facilement interprétable mais dont le principe repose sur l'existence d'une relation linéaire entre les variables. Les deux méthodes sont prisés et sont couramment utilisés en apprentissage supervisé pour leurs bonnes performances et leur capacité à se généraliser même face aux problèmes les plus complexes mais restent assez difficiles à interpréter.

4.1 Les modèles de régression linéaire

Avant d'introduire la régression LASSO, il est important d'introduire d'aborder le concept des régressions linéaires.
Les modèles de régression font partie des modèles les plus simples et les plus intuitifs que ce soit en apprentissage statistique ou en Machine Learning. Il existe de nombreux types de régressions linéaires et après de nombreux tests, nous avons décidé de présenter uniquement les résultats de la régression LASSO car ceux-ci étaient meilleurs que les autres. Toutes fois, afin de comprendre le fonctionnement de la régression LASSO, il est important d'aborder la régression Linéaire simple et la régression Ridge.

4.1.1 Régression linéaire simple

Un modèle de régression linéaire est un modèle qui suppose que notre variable cible est une combinaison linéaire de nos différentes variables à expliquer. Le principe de la régression linéaire repose donc sur la supposition d'une relation linéaire entre nos variables.
La régression linéaire simple ou régression linéaire est affine est la plus basique. Elle consiste à estimer notre variable à expliquer Y comme à l'aide d'une fonction affine de X. En terme mathématiques, cela s'écrit :

$$f(X_i) = \beta_0 + \beta_1 X_i^1 + ... + \beta_p X_i^p + \varepsilon_i,\ 1 \leq i \leq n$$

Où,

— $(\beta_i)_{1 \leq i \leq n}$ est le vecteur réel des paramètres de la régression qu'il faudra estimer
— Les ε_i sont des variables aléatoires qui vérifient :
  1. $\mathbb{E}[\varepsilon_i] = 0$
  2. $Cov(\varepsilon_i, \varepsilon_j) = 0$
  3. $\forall i \neq j,\ Var(\varepsilon_i) = \sigma^2$

On peut également l'écrire sous forme matricielle :

$$f(X) = X\beta + \varepsilon$$

Où,

$$X = \begin{pmatrix} X_1^0 & X_1^1 & . & . & X_1^p \\ X_2^0 & X_2^1 & . & . & X_2^p \\ . & . & . & . & . \\ . & . & . & . & . \\ X_n^0 & X_n^1 & . & . & X_n^p \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_p \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ . \\ . \\ \varepsilon_p \end{pmatrix}$$

$$X^0 = \begin{pmatrix} X_1^0 \\ X_2^0 \\ . \\ . \\ X_n^0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ . \\ . \\ 1 \end{pmatrix}, X^j = \begin{pmatrix} X_1^j \\ X_2^j \\ . \\ . \\ X_n^j \end{pmatrix}, (j = 1, ..., p)$$

Le but de tout modèle étant de faire une prédiction qui se rapproche au maximum de la variable théorique, trouver les paramètres revient à minimiser les ε_i. En régression linéaire, cela se traduit par la minimisation de l'erreur quadratique. Le vecteur β qui correspond aux paramètres de la régression est celui qui respecte :

$$\min_{\beta} \|X\beta - Y\|_2^2$$

En d'autres termes, l'estimateur β^* minimise :

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i^1 - ... - \beta_p X_i^p)^2$$

La technique des moindres carrées ordinaires est utilisée pour trouver la solution. L'estimateur obtenu est le suivant :

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Si la régression linéaire est un modèle simple et facilement compréhensible, elle présente de nombreuses pénalités.
Premièrement, elle suppose que la valeur à expliquer Y suit une fonction affine, ce qui ne sera pas souvent le cas dans la réalité. Cela crée déjà un certain biais.
Ensuite, il existe une hypothèse non colinéarité des variables explicatives. En effet, si le modèle comporte des variables colinéaires, alors la matrice X ne pourrait pas être inversible et il n'y aurait alors pas de solution pour l'équation permettant de trouver β.

Afin de résoudre le dernier problème, E.Hoerl et Kennard proposent une solution dans leur ouvrage « Ridge régression : biaised estimation for nonorthogonal problems, Technometrics, Vol. 1 Février 1970 ». C'est la naissance la Régression Ridge.

4.1.2 Régression Ridge

La régression Ridge est basée sur le même principe que la Régression linéaire et vient pallier au problème concernant la colinéarité des variables. C'est une régression linéaire avec une contrainte quadratique sur les coefficients. C'est utile lorsque les variables sont très corrélées, ce qui fausse souvent la résolution numérique. La solution peut s'exprimer de façon exacte. La régression Ridge s'écrit :

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i^1 - ... - \beta_p X_i^p)^2 + \alpha \sum_{i=1}^{n} \beta^2$$

Ceci se traduit aussi par :

$$\min_{\beta} \|X\beta - Y\|_2^2 + \lambda \|\beta\|_2^2$$

α est appelé coefficient de pénalité sur la colinéarité des variables. L'estimateur de Ridge est $\beta_{Ridge} = (X'X + \alpha I_{p+1})^{-1} X'Y$. On peut donc choisir α tel que X soit inversible.

Le coefficient de pénalité contrôle la quantité de rétrécissement : plus sa valeur est grande, plus le rétrécissement est important et donc plus les coefficients sont robustes à la colinéarité.

4.1.3 Régression LASSO

Le LASSO (Least Absolute Shrinkage and Selection Operator) est un modèle linéaire qui estime des coefficients épars. Il est fortement utile dans les situations où il faut gérer une multitude de variables potentiellement liées. En effet, ce modèle a la capacité de mettre à zéro les coefficients de certaines variables, ce qui réduit le nombres de variables à utiliser pour le modèle.

En considérant le même problème qui a mené à la régression Rdige, Robert Tibshirani dans un article publié en 1996 intitulé «Regression shrinkage and selection via the LASSO», choisit comme contrainte de minimiser les coefficients B_j lors de la modélisation.
Le programme de minimisation devient alors :

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i^1 - ... - \beta_p X_i^p)^2 + \lambda \sum_{i=1}^{n} |\beta_i|$$

Ceci se traduit aussi par :

$$\min_{\beta} \|X\beta - Y\|_2^2 + \lambda \|\beta\|_1$$

Avec $\lambda(\lambda \geq 0)$ le coefficient de pénalité de la régression.

Minimiser cette expression revient à minimiser l'expression de régression linéaire simple avec la contrainte : $\lambda \sum_{j=1}^{p} |\beta_j| \leq s$, s un réel

.

Remarques :

— Lorsque $\lambda = 0$, l'équation de LASSO devient l'équation de régression linéaire simple

— Lorsque λ augmente, le biais a tendance à augmenter et la variance à diminuer

— Lorsque $\lambda \to \infty,\ \beta_{Lasso} \to 0$

Par rapport à la régression Ridge la régression LASSO a pour avantage de mettre à 0 les effets peu importants, donc le modèle sélectionné aura un nombre de variables $d < p$, où p représente le nombre de variables du modèle initial. Le modèle choisit donc lui-même les variables qui sont importantes pour lui. En présence de variables explicatives corrélées, le LASSO en choisit arbitrairement une et met les autres à 0.

Nous pouvons maintenant passer aux modèles XGBoost et random forest mais avant cela, nous allons introduire les modèles d'arbre de décisions, qui n'est pas utilisé dans notre démarche mais qui aidera à mieux comprendre les modèles que nous avons choisis. En effet les modèles d'arbres de décisions sont la racine de plusieurs méthodes ensemblistes.

4.2 Introduction aux arbres de décision

Les arbres de décision ou algorithme CART (Classification And Regression Trees) ont été en mis place en 1984 par Leo Brieman et constituent désormais la base des méthodes ensemblistes. L'algorithme CART permet d'expliquer aussi bien une variable qualitative (on parle alors d'arbre de régression) que qualitative (on parle d'arbre de discrimination).

Les arbres de décisions sont des algorithmes simples et assez performants, en plus d'être non linéaires et non paramétriques. Contrairement aux algorithmes complexes, ils ont le grand avantage de pouvoir se présenter sous une forme graphique simple à visualiser permettant de bien comprendre les résultats. Malheureusement, ces derniers sont très dépendants des données d'apprentissage ce qui les rend très instables. Le modèle est donc propice à faire du surapprentissage et à produire des résultats avec une forte variance.

- **Les méthodes ensemblistes**

Les méthodes ensemblistes ont donc été développés pour offrir une meilleure stabilité aux algorithmes CART et résoudre ce problème de variance, principal défaut de l'arbre de décision. Encore appelées méthodes d'agrégation, les méthodes ensemblistes consistent à construire un grand nombre d'arbres de décisions et à les agréger entre eux en utilisant une méthode spécifique. Ainsi l'estimateur final aura une variance considérablement réduite et sera plus consistant.

Les techniques utilisées pour construire et agréger les arbres font la particularité de chaque modèle ensembliste. Les deux techniques les plus connus et également celles qui nous serviront dans notre démarche sont le Bagging et le boosting.
Le Bagging consiste en l'assemblage des arbres construits aléatoirement et indépendamment les uns des autres ;cette technique donne naissance au random forest.
Dans le cas du boosting, chaque arbre construit a pour but de corriger le précédent. Il y'a donc une forte dépendance entre chaque arbre ; cette technique donne naissance au XGBoost.

- **Principe de l'algorithme CART**

L'algorithme CART est fondé sur des arbres et cherche à diviser localement les données en plusieurs petits segments en fonction de différentes valeurs et combinaisons de prédicteurs. CART identifie les divisions les plus performantes, puis répète ce processus régulièrement jusqu'à obtenir le résultat idéal. Il en résulte un arbre de décision représenté par une série de divisions binaires débouchant sur des nœuds terminaux qui peuvent être décrits par un ensemble de règles spécifiques. Il se démarque donc des autres solutions d'analyse prédictive grâce à sa méthodologie unique et pertinente dont les maîtres mots sont l'automatisation, la facilité d'utilisation, la performance et la précision.

Un arbre binaire est une construction hiérarchique de forme "triangulaire en escaliers" constitué de plusieurs éléments : l'élément fondateur est au sommet de la construction ; il est appelé racine, les traits qui partent en descendant de cette racine sont appelés branches, elles joignent des éléments appelés nœuds. De chaque nœud, partent 0 ou 2 branches joignant alors d'autres nœuds, et ainsi de suite. Un nœud dont part 2 branches est dit coupé. Un nœud dont ne part aucune branche est appelé feuille. Ainsi, un arbre se parcourt de la racine aux feuilles (donc de haut en bas).

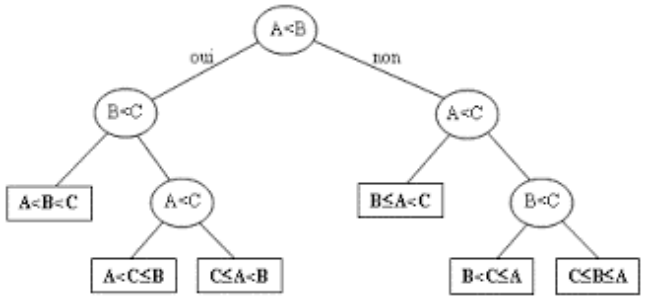

FIGURE 2.4 – Exemple d'arbre de décision

- **Fonctionnement de l'algorithme CART**

Les arbres de décision sont basés sur un découpage de l'espace engendré par les variables explicatives en plusieurs nœuds. Pour prédire une valeur y d'un portefeuille dont on connaît les caractéristiques $(x_1, x_2, ..x_k)$, on procède en fonction des étapes suivantes :

— En partant de la racine à chaque nœud, on vérifie si la condition de division est vérifiée ou pas. Si elle est vérifiée, on se dirige vers la branche associée 'Oui'. Sinon, on se dirige vers la branche associée 'Non'

— On réitère le processus à chaque nœud jusqu'à se retrouver à un nœud ne donnant naissance à aucune branche, c'est-à-dire dans une feuille

— A ce moment, la valeur possible de y sera alors la moyenne des valeurs des y de cette feuille.

La construction d'un modèle CART reposera alors sur les différents aspects suivants : un critère de division de nœud pour indiquer de quelle façon scinder les nœuds, une règle d'arrêt permettant de déterminer quand un nœud est terminal (on parle aussi de feuille), un critère d'affectation ainsi qu'un critère d'homogénéité.

L'image suivante résume le fonctionnement de l'algorithme :

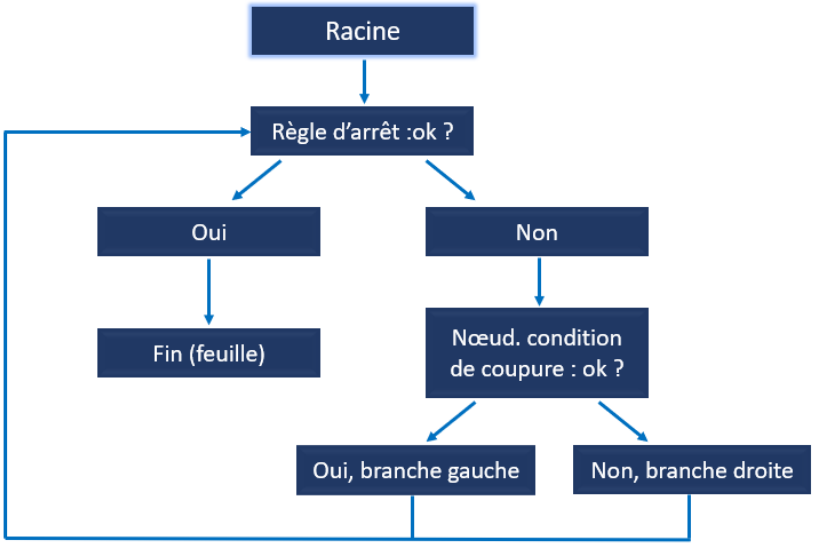

FIGURE 2.5 – Mise en place d'un arbre de décision

- **Critère de division**

Soit $j \in \{1, ..., k\}$ et $c \in \mathbb{R}$. Pour un nœud donné, on pose :

— $\bar{Y}_{j,c,gauche}$ la moyenne des valeurs de Y pour les individus qui vérifient $X_j < c$

— $\bar{Y}_{j,c,droit}$ la moyenne des valeurs de Y pour les individus qui vérifient $X_j \geq c$

— $\text{SS}_{gauche}(j, c)$ la somme des carrés des écarts entre les valeurs de Y et $\bar{Y}_{j,c,gauche}$ pour les individus vérifiant $X_j < c$

— $\text{SS}_{droit}(j, c)$ la somme des carrés des écarts entre les valeurs de Y et $\bar{Y}_{j,c,droit}$ pour les individus vérifiant $X_j \geq c$

Pour un nœud donné, l'erreur commise en séparant les observations selon les critères $X_j < c$ où $X_j \geq c$ est

$$E(j, c) = SS_{gauche}(j, c) + SS_{droit(j,c)}$$

Le but de chaque division étant de minimiser cette erreur, il faut trouver les valeurs j_* et c_* tels que :

$$\forall j \in 1, ..., k \; et \; \forall c \in \mathbb{R}, \quad E(j_*, c_*) \leq E(j, c)$$

La variable X_{j_*} sera alors la variable utilisée pour couper le nœud et c_* sera le seuil de découpage. Les observations seront donc séparées en deux groupes suivant qu'elles vérifient $X_j < c_*$ ou $X_j \geq c_*$. Le choix de la variable de découpage à la racine nous indiquera un facteur le plus influent sur Y.

- **Règle d'arrêt**

Une règle d'arrêt nous permettra de savoir à quel moment cesser d'appliquer le critère de division sur un nœud. Il existe plusieurs critères d'arrêt qu'on peut fixer à l'algorithme :

— La profondeur de l'arbre

— Le nombre minimum d'individus présents à l'étape d'un nœud pour envisager une coupure (ce nombre doit être supérieur à 0)

— Le nombre minimum d'individus présents à l'étape d'un nœud fils engendré par un nœud père (en dessous de ce nombre, le nœud père devient alors une feuille)

— La valeur d'un paramètre de complexité , proportionnel à la taille de l'arbre

Remarque : Si aucun critère d'arrêt n'est choisi, l'algorithme fera des séparations de nœuds jusqu'à avoir des feuilles ne contenant qu'une observation chacune. On parle alors d'arbre maximal.

- **Élagage**

Une fois l'arbre construit à partir des deux critères ci haut, si l'arbre est trop profond (ce qui arrive souvent lorsqu'aucun critère d'arrêt n'est choisi), le modèle procède à un élagage qui vise à supprimer les branches des feuilles vers la racine. Le nombre de feuilles est alors réduit et on obtient plus d'observations dans les feuilles touchées, ce qui aide à réduire la variance des résultats obtenus.

L'objectif de l'élagage est d'obtenir un bon compromis entre la complexité α et la qualité de prédiction. Si α est nul, l'arbre obtenu minimise l'erreur de complexité. Par contre un α élevé favorisera des arbres avec peu de subdivisions et donc peu de feuilles. Ce qui va pénaliser le biais au profit de la variance. Tandis que la réduction des branches pénalisera la variance au profit du biais. La recherche du d'un bon équilibre quant au nombre de branches à supprimer nous ramène sur fameux dilemme biais-variance.

Nous allons à présent aborder les deux algorithmes d'apprentissage sélectionnés pour notre démarche. Il s'agit du random forest et du XGBoost, tous les deux des modèles ensemblistes créés sur la base des arbres de décision. Tandis que l'algorithme de random forest consistera en la création d'arbres au biais faibles pour les agréger en réduisant leur variance, le XGBoost agrégera plusieurs arbres à variance faible en cherchant à réduire le biais.

4.3 Le random forest

L'algorithme random forest ou forêts aléatoires a été créé en 1995 par HO, puis formellement proposé par les scientifiques Adele Cutler et Leo Brieman en 2001. C'est un algorithme bien connu dans le domaine du Machine Learning pour son efficacité en prédiction et qui peut être utilisé dans des tâches de régression ou de classification.

Il s'agit d'un modèle ensembliste, c'est-à-dire qu'il « met ensemble » ou agrège plusieurs arbres de décision de façon intelligente pour former un modèle plus complexe et plus performant. De nombreux articles de Machine de Learning effectuant des comparaisons entre divers méthodes d'apprentissage abordent le random forest, qui est souvent désigné comme vainqueur. Son efficacité lui a permis d'être utilisé dans de nombreux domaines comme le marketing téléphonique pour prédire le comportement de clients ou encore la finance pour la gestion de risques.

La mise en place de l'algorithme random forest, repose essentiellement sur deux principes : Le Bagging et le feature sampling.

- **Le Bagging**

Le Bagging (diminutif de boostrap aggregation) est une technique ensembliste qui permet de générer une multitude d'arbres par boostrap puis de les agréger, afin de corriger l'instabilité des modèles CART.
Le Bagging consiste à créer de façon aléatoire un nombre B d'arbres de décisions suivant les étapes suivantes :

1. Découpage la base d'apprentissage de façon aléatoire en B sous échantillons de taille m. C'est-à-dire qu'on effectue n tirages aléatoires et avec remise dans la base d'apprentissage pour créer un sous échantillon de taille m. Ensuite on réitère l'opération jusqu'à obtenir B sous échantillons de taille n. On notera les sous échantillons (X_b, Y_b), avec b un entier compris entre 1 et n . Il est donc possible d'avoir plusieurs observations identiques dans un sous échantillon donné. Ce processus s'appelle de Booststrapping.
2. Un modèle CART est entraîné sur chaque sous ensemble (X_b, Y_b). On construira donc au total un nombre B de modèles CART différents, puisqu'ils auront été entraînés sur des bases de données différentes.
3. Le résultat final est obtenu en agrégeant les résultats de chaque arbre en faisant la moyenne dans le cas d'une régression et par système de vote(On choisit la classe la majoritaire) dans le cas d'une classification.

Cette technique permet de réduire la variance et le risque de surapprentissage, bien qu'elle conduise parfois à un biais élevé. L'image ci-dessous résume le procédé du Bagging :

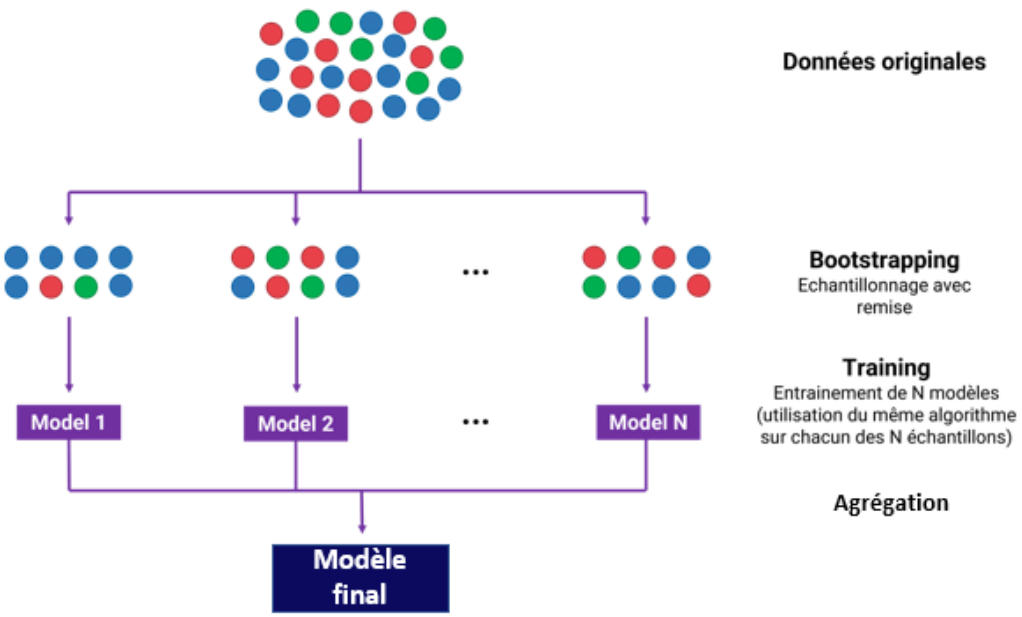

FIGURE 2.6 – Illustration du Bagging

- **Le feature sampling**

Le feature sampling qui peut littéralement être traduit par « échantillonnage des variables » est une technique assez similaire au Bagging mais qui propose une sélection aléatoire des variables explicatives et pas des observations. Le feature sampling consiste donc à créer des modèles basés sur des

sous échantillons avec un nombre restreint de variables explicatives choisies de façon aléatoires. Le modèle dépend donc fortement de la pertinence des variables explicatives. En effet, si des arbres sont créés avec uniquement des variables aléatoires ne contenant aucune information utile, cela engendrera un mauvais résultat qui réduira la précision du modèle final.

Le random forest est donc un mélange de Bagging et de feature sampling. La sélection aléatoire des observations combinée à la sélection aléatoires des variables explicative permet de garantir l'indépendance des arbres et donc de réduire la variance.

En effet, la variance de la moyenne (c'est-à-dire la variance du modèle final) de B variables indépendantes, identiquement distribuées, chacune de variance σ^2, est σ^2/B. Si ces variables sont identiquement distribuées avec une corrélation ρ des variables deux à deux, alors la variance de la moyenne devient :

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

L'avantage du Bagging fait que le deuxième terme sera limité si ρ est grand. Et le feature sampling limite le premier terme si ρ est petit. Enfin, plus B est grand, plus le second terme sera petit.

Chaque sous arbre du random forest pourra donc traiter une partie du problème. L'agrégation de tous ces arbres permettra d'avoir une vision plus générale comme illustré dans l'image ci-dessous.

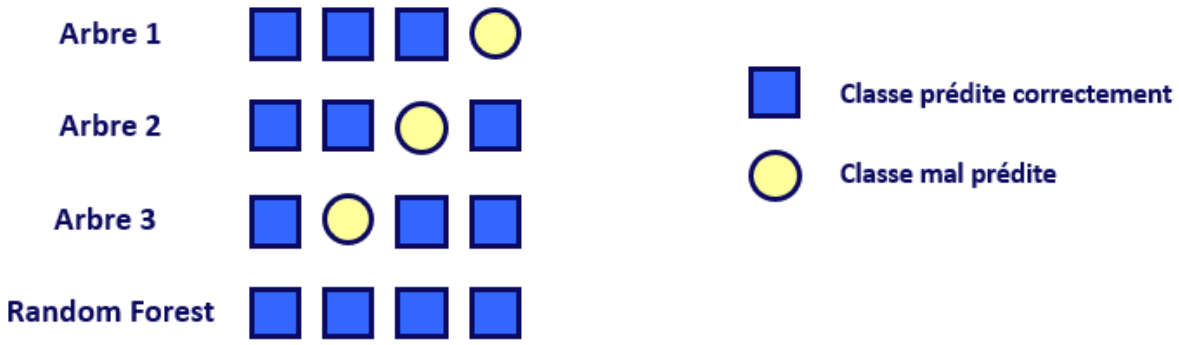

FIGURE 2.7 – Illustration d'une agrégation d'arbres

Il faudra cependant veiller à fournir aux modèles des variables explicatives appropriées pour éviter de mauvais résultats. Heureusement, le random forest offre la possibilité de mesurer l'importance de chacune des variables explicatives dans la prédiction de la variable cible. Ceci jouera un grand rôle dans la sélection des variables pour la suite.

- **Implémentation du modèle et paramétrisation**

Le modèle est implémenté sous python grâce à la bibliothèque Scikit Learn. La paramétrisation du modèle permet son optimisation et donc de meilleurs résultats. Par exemple, le nombre d'arbres que avons déjà mentionné plus haut joue un grand rôle dans la construction du modèle. Il s'agit en effet

d'un paramètre à choisir lors de l'implémentation du modèle. Il existe plusieurs autres paramètres, mais voici les plus importants :

— $n\_estimators$: le nombre d'arbres à entraîner

— max_depth : la profondeur maximale de chaque arbre

— min_sample_split : le nombre minimal d'observations par feuille, il s'agit d'un critère d'arrêt

— $max\_features$: le nombre de variables tirées aléatoirement pour chaque arbre lors du feature sampling. Par défaut il vaut $\sqrt{k}$ dans le cas d'une classification et $\frac{k}{3}$ dans le cas d'une régression

— $criterion$: le critère utilisé pour couper les feuilles de chaque arbre en cours de construction, par défaut ce critère est la MSE

Les autres paramètres sont disponibles sur le site de scikit-learn et dans l'aide Python. La paramétrisation est faite grâce au Gridsearch et à la validation croisée, tous deux implémentés via scikit-learn et qui nous permettent de tester et de juger différents paramétrisations possibles pour le modèle. Nous aborderons ces deux notions dans la suite.

En résumé, Le principe du random forest réside en la construction de plusieurs arbres CART indépendants qui seront ensuite agrégés dans le but de réduire l'erreur de prédiction et la variance. C'est un algorithme performant mais dont l'interprétabilité est assez difficile à cause de l'effet « boite noire » créé par l'agrégation d'une multitude d'arbres.

4.4 Le XGBoost

Tout comme le random forest, le modèle XGBoost (Extreme Gradient boosting) repose également sur une agrégation de plusieurs arbres CART. Mais contrairement au random forest, la construction des arbres se fait de manière séquentielle et non indépendante. Sa construction repose sur les techniques de Boosting et du Gradient Descent.

4.4.1 Le Boosting

Le Boosting est une méthode ensembliste qui consiste à assembler plusieurs modèles « faibles » pour former un modèle « fort ». Cette technique consiste en la construction de plusieurs modèles de façon séquentielle, de manière que chaque modèle puisse réduire l'erreur de prédiction commise par le modèle précédent. Contrairement au Bagging qui consiste à l'assemblage de modèle de façon indépendante, le boosting va produire des modèles qui sont très dépendants les uns des autres.

Le Boosting s'implémente selon les étapes suivantes :

1. **Construction d'un premier modèle** : A partir d'un algorithme choisi, le boosting va construire un tout premier modèle. Cet algorithme peut être une régression linéaire ou encore un algorithme CART . Cet algorithme est ensuite entraîné sur les données d'apprentissage et toutes les observations ont le même poids.
2. **Construction du second modèle** : A l'aide de la fonction d'erreur choisie, le premier modèle est évalué et lorsqu'une observation est mal prédite, celle-ci voit son poids augmenter. Le second modèle est donc entraîné à l'aide des données pondérées obtenues. Son but étant de corriger les erreurs du premier modèle c'est-à-dire de se concentrer davantage sur les observation ayant un poids élevé.

3. **Modèle final** : La procédure précédente se poursuit et des modèles sont ajoutés jusqu'à ce que l'ensemble complet des données d'apprentissage soit prédites correctement ou que le nombre maximal de modèles soit ajouté. Les prédictions du dernier modèle construit représentent les prédictions globales fournies par les anciens modèles. C'est le modèle final obtenu.

On peut donc déjà détecter deux paramètres significatifs pour l'implémentation d'un modèle de Boosting : la fonction d'erreur permettant de juger de la qualité de prédiction d'une observation et le nombre de modèle maximal qui est aussi un critère d'arrêt pour l'algorithme.
Il existe différents algorithmes de Boosting, qui diffèrent selon le modèle basique choisi, la fonction d'erreur, la façon de pondérer les observations, etc...

Nous nous intéressons uniquement aux cas où les modèles basiques choisis sont des arbres de décision, comme dans le cas du XGBoost.

4.4.2 Gradient Descent

Nous allons maintenant présenter une technique qui découle également du Boosting, le Gradient Boosting. Cette méthode consiste à agréger les modèles à l'aide de la méthode du Gradient Descent. Elle peut être utilisée pour plusieurs types de modèles mais nous intéresserons uniquement au cas des arbres de régressions.

Cette méthode forme des modèles de régression en ajustant par moindres carrés, à chaque itération, une fonction appelée base learner à des pseudo-résidus. Ces derniers forment le gradient d'une fonction de perte qui doit être minimisée, ce gradient étant approché par un arbre de régression. À chaque étape, le modèle agrégé apparaît comme un pas vers la solution optimale, ce pas étant fait dans la direction du gradient de la fonction de perte.

La fonction des moindres carrées se définit comme suit :

$$L(Y, f(X)) = \sum_i \frac{(Y_i - f(X_i))^2}{2}$$

Les étapes du Gradient Boosting sont les suivantes :

- **Initialisation**

On commence par initialiser le processus en construisant un premier arbre f_0 qui aura pour objectif de prédire Y en fonction de X. Pour toute observation i, l'erreur commise par la fonction f_0 sera

$$\varepsilon_i = Y_i - f_0(X_i) = -\frac{\partial L(Y_i, f_0(X))}{\partial(f_0(X))}$$

En effet,

$$\frac{\partial L(Y_i, f_0(X))}{\partial(f_0(X))} = \frac{\partial(Y_i - f_0(X))^2}{2\partial(f_0(X))}$$

$$= -2 \times \frac{(Y_i - f_0(X))}{2}, \text{ on rappelle que } \frac{\partial(a-S)^2}{\partial S} = 2(a-S)$$

$$= -(Y_i - f_0(X))$$

- **Réduction de l'erreur**

Ensuite, on construit un arbre f_1 qui aura pour but de réduire l'erreur commise par f_0. L'intérêt est de capter les observations qui n'ont pas pu être prédites par le modèle précédent. Pour cela, un nouveau modèle faible j est créer sur les données (X, r) et on détermine le poids associé à ce modèle faible comme :

$$\gamma_m = argmin_\gamma \sum_{i=1}^{n} L(Y_i, f_{m-1}(X_i) + \gamma h_m(X_i))$$

Le nouveau modèle s'écrit alors comme :

$$f_m(X) = f_{m-1}(X) + \eta\gamma_m h_m(X)$$

Avec η un paramètre appelé taux d'apprentissage. Il permet d'éviter le surapprentissage. C'est un paramètre à choisir lors de l'implémentation du modèle. Le nouveau résidu à minimiser est désormais : $\varepsilon_2 = Y_i - f_2(Xi)$

- **Itération**

— On réitère l'opération jusqu'à ce que les résidus soient nuls ou que le nombre d'arbres maximal soit atteint. A l'étape m, le modèle peut donc s'écrire comme :

$$f_m(X) = f_{m-1}(X) + \eta\gamma_m h_m(X)$$

4.4.3 Le XGBoost

En suivant le même principe que le Gradient Descent, le XGBoost consiste à construire de façon successive plusieurs arbres, de façon que chaque arbre minimise l'erreur commise par l'arbre précédent.

Le XGBoost utilise comme fonction de coût la MSE :

$$\Omega(Y, f(X)) = \sum_i L(Y - f(X))^2$$

Le XGBoost est en fait une version particulière de l'algorithme de Gradient Boost. En effet, il s'agit d'un assemblage de "weak learners" qui prédisent les résidus, et corrigent les erreurs des "weak learners" précédents. La particularité d'XGBoost réside dans le type de "weak learner" utilisés, qui sont des arbres décisionnels.

Les arbres qui ne sont pas assez bons sont élagués, c'est à dire qu'on leur coupe des branches, jusqu'à ce qu'ils soient suffisamment performant. Sinon, ils sont complètement supprimés. Cette méthode est appelée le "pruning" (élagage). Ainsi, le XGBoost s'assure de ne conserver que de bons weak learners.

- **Implémentation du modèle et paramétrisation**

Dans cette étude, nous utiliserons l'algorithme « XGBRegressor » du package « XGBoost » de Python. Ci-après les principaux paramètres :

— *n_estimators* : le nombre d'arbres

— *learning_rate* : correspond au paramètre η présenté précédemment

— λ : le niveau de baisse minimale de la fonction objectif pour faire une subdivision supplémentaire

— *max_depth* : profondeur maximale d'un arbre

— *min_child_weight* : nombre minimum d'observations par feuille

5 Validation croisée et Gridsearch

Lors de l'application d'un modèle de Machine Learning, le choix des paramètres est une étape cruciale pour l'obtention de meilleurs résultats. Par exemple pour un random forest, on doit choisir le nombre d'arbres à créer et le nombre de variables à utiliser à chaque division d'un nœud. Si on paramètre à la main, cela peut vite s'avérer très coûteux en temps et pas forcément très intéressant. C'est là que le Gridsearch intervient.

Le GridSearch est une méthode d'optimisation (hyperparameter optimization) qui va nous permettre de tester une série de paramètres et de comparer les performances pour en déduire le meilleur paramétrage. Il existe plusieurs manières de tester les paramètres d'un modèle et le GridSearch est une des méthodes les plus simples. Pour chaque paramètre, on détermine un ensemble de valeurs que l'on souhaite tester. Par exemple, dans le cas d'un random forest on pourrait tester le nombre d'arbres ou le nombre de nœuds.
En plus de cela, le Gridsearch utilise une méthode de validation croisée : Le k-fold.

Une fois que chaque combinaison de paramètres a pu être testée et évaluée, il ne reste plus qu'à comparer les performances pour choisir la meilleure combinaison. Le Gridsearch offre la possibilité de choisir une unité de mesure de performance entre le R2 et le MSE. Dans notre cas, nous avons choisi le MSE. Le meilleur modèle sera celui avec la MSE la plus faible.

La validation croisée est une technique couramment utilisée en Machine Learning pour évaluer les performances d'un modèle. Elle consiste à diviser l'ensemble des données en plusieurs sous-ensembles, d'entraîner le modèle sur une partie des données et de tester sa performance sur les autres parties (appelée base de validation). Ceci permet d'obtenir une estimation de la précision du modèle sur l'ensemble des données.

Le principe de la validation croisée est de répéter cette opération pour chaque combinaison de sous-ensembles, en faisant tourner chaque sous-ensemble à la fois pour l'entraînement et le test. Cela permet d'éviter le surapprentissage du modèle sur un ensemble de données spécifique et offre une évaluation fiable de sa performance.
Il existe une multitude de techniques de validation croisée, qui varient les unes des autres en fonction de la manière dont la base est divisée en sous ensembles. Nous ne présenterons que le K-fold qui a été utilisé lors de nos travaux.

- **La méthode des k-Fold**

Il s'agit d'une méthode de validation croisée simple à comprendre et assez populaire. Elle permet de garantir que toutes les observations puissent apparaître dans la base d'entraînement et dans la

base de validation. C'est donc une bonne option dans le cas où on ne dispose pas de suffisamment de données, ou lorsqu'on juge que chaque ligne d'observation pourrait avoir un impact sur le modèle. Le processus de cette méthode se décompose comme suit :

1. On sépare l'ensemble des données des données de façons aléatoires en k échantillons. K ne doit pas être trop grand ni trop haut, la valeur choisie par défaut est $k = 5$
2. On apprend ensuite le modèle sur les $k-1$ premiers échantillons et on le valide sur les k-échantillon restant. On enregistre l'erreur et le score obtenu sur ce $k - iéme$ échantillon
3. Le processus est répété k fois, jusqu'à ce que chaque échantillon serve de base de validation
4. On calcule le score final en faisant la moyenne de tous les scores enregistrés

L'image ci-dessous issue du représente parfaitement le procédé du K-fold lorsque k=5

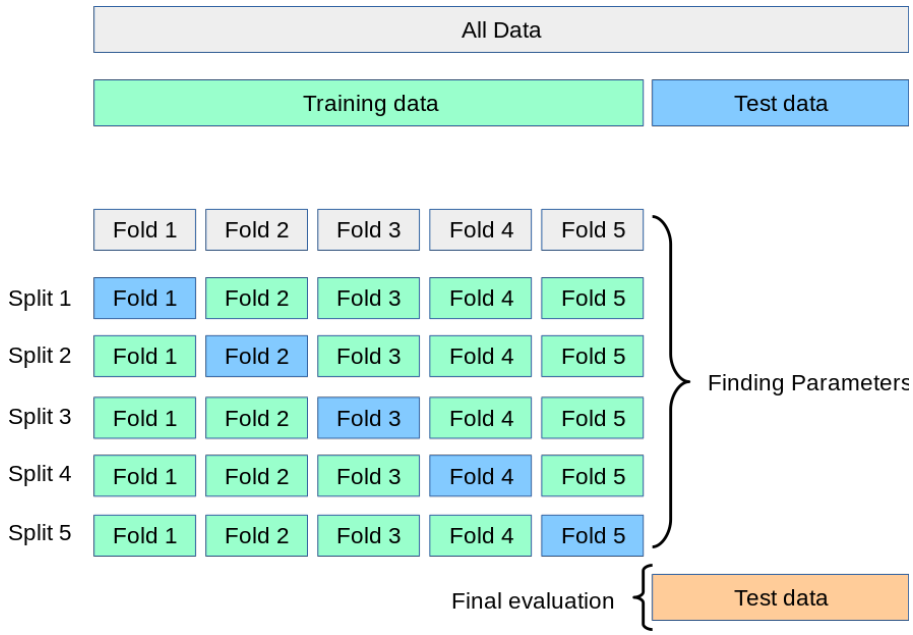

FIGURE 2.8 – Validation croisée K-fold

Comme on peut le voir, sur cette image, la base d'apprentissage est séparée en 5 échantillons. A chaque itération, un modèle sera entraîné sur les ensembles en verts et testé sur l'ensemble de couleur bleue (base de validation). Le résultat final du modèle en question sera obtenu en faisant l'agrégation des résultats obtenus sur chacun des 5 ensembles de validation.

Cette technique pendant le Gridsearch afin de déterminer quel combinaison de paramètre permettra une meilleure généralisation du modèle à de nouvelles données.

Lorsque $k = n$, c'est-à-dire que le nombre de division est égal au nombre d'échantillons disponibles, on parle de Leave One Out (LOO). C'est également une méthode de cross validation assez répandue. L'algorithme est alors entraîné sur toute la base d'apprentissage sauf une observation puis testé sur cette unique observation.

Chapitre 3

Description du modèle épargne

1 Présentation d'un produit d'épargne

1.1 Définition d'un contrat d'assurance

L'Article 1101 du Code Civil définit un contrat comme suit : « Le contrat est une convention par laquelle une ou plusieurs personnes s'obligent, envers une ou plusieurs autres, à donner, à faire ou à ne pas faire quelque chose. » et le contrat d'assurance respecte le cadre de cette définition générale.

Un contrat d'assurance est « un contrat par lequel le souscripteur se fait promettre par un assureur, pour son compte ou celui d'un tiers (l'assuré), une prestation généralement pécuniaire en cas de réalisation d'un risque, moyennant le paiement d'une prime ou cotisation. »

Un risque est un évènement incertain se produisant indépendamment de la volonté de l'assuré et du souscripteur. En d'autres termes, un risque représente la probabilité qu'un dommage survienne de manière aléatoire. Le souscripteur s'assure pour se protéger dans le cas où cet évènement se produirait. On parle du caractère aléatoire du contrat d'assurance. L'incertitude sur la réalisation ou non d'un évènement, ou encore sur la date de réalisation d'un évènement est indispensable pour la mise en place d'un contrat d'assurance. Sans aléa, il n'y a pas d'assurance.

Il existe plusieurs types de contrats d'assurance mais dans le cadre de cette étude, nous n'aborderons que le cadre de l'assurance vie.
L'assurance vie est « un contrat par lequel en échange d'une prime, l'assureur s'engage à verser au souscripteur ou au tiers par lui désigné, une somme déterminée (capital ou rente) en cas de décès de la personne assurée ou de sa survie à une époque déterminée. » Étant donné que le décès d'une personne est un évènement certain qui finira par arriver, l'aléa permettant de mettre en place le contrat repose donc sur la date de décès de la personne assurée.

Cette définition de l'assurance vie nous emmène à distinguer deux cas :

— L'assurance en cas de décès ou assurance décès qui garantit à l'assuré le versement d'un capital ou d'une rente en cas de décès du souscripteur avant le terme du contrat

— L'assurance en cas de vie ou assurance vie pure, qui garantit à l'assuré le versement d'un capital défini s'il est toujours en vie à l'échéance du contrat. La date d'échéance et le montant du capital sont définis de façon claire dans le contrat

L'assurance en cas de décès constitue une garantie pour les proches du souscripteur tandis que l'assurance en cas de vie est davantage utilisée comme placement, étant donné que l'assuré peut lui-même être bénéficiaire. Cependant, il existe des contrats dits mixtes qui combinent les deux garanties.

Une fois de plus, nous allons nous restreindre seulement au sujet qui nous intéresse, c'est-à-dire l'assurance en cas de vie, souvent utilisée pour épargner de l'argent. En effet, l'assurance vie permettra au souscripteur d'épargner de l'argent pour sa retraite ou pour financer un projet à moyen terme. Il s'agit du premier moyen d'épargne en France, ce qui est dû en grande partie aux avantages fiscaux qu'il présente.

1.2 Présentation des contrats d'épargne

On distingue trois grand types d'épargne, qui diffèrent principalement par la façon dont sont investies les primes de l'assuré. Ce dernier choisit le contrat qui lui correspond le mieux en fonction de la durée envisagée de ses placements et de son appétence au risque.

- **Les supports en euros**

Les supports en euros représentent le fond d'épargne le plus prudent de tous. En effet, le capital investi est garanti à tout moment et les intérêts de l'année sont définitivement acquis. Les versements sont investis sur des produits sans risque, tels que des obligations d'État, et revalorisés chaque année. Ces intérêts versés chaque année à l'assuré dépendent d'un taux minimal garanti (TMG) et d'une clause de participation aux bénéfices (PB) définis lors de la souscription du contrat. La clause de PB permet de définir la part minimale à verser à l'assuré sur le rendement des investissements de l'assureur. Nous détaillerons cette partie plus tard.

Les supports en euros sont donc destinés aux personnes souhaitant prendre le moins de risque possible tout en ayant la possibilité d'accès à leur capital à tout moment. En contrepartie, les rendements obtenus sont généralement faibles. Nous avons effectué nos travaux sur un environnement fictif pour lesquels les taux sont très bas, et le TMG fixé à 0% en France. Il est donc possible qu'en cas de mauvais rendement des investissements, que l'assuré ne reçoive aucun bénéfice annuel sur son épargne.

- **Les supports en unités de compte**

Les supports en Unité de Compte (UC) permettent d'investir l'épargne sur des supports très variés tels que les marchés financiers, l'immobilier, les sociétés, etc... L'assuré accepte de prendre un risque plus élevé, contre la possibilité d'un rendement plus intéressant. Une UC représente un support d'investissement tel que la part d'une action et n'offre donc pas de garantie sur la valeur du rendement. Dans ce contrat, l'assureur ne garantit que le nombre d'UC et pas le capital initial de l'assuré en cas de baisse de l'UC. L'assuré accepte donc le risque de perdre une partie de son capital qui dépend des fluctuations des marchés.

Ces contrats sont destinés aux personnes avec une certaine appétence au risque et qui de plus ont la possibilité de laisser mûrir leur épargne suffisamment longtemps. En effet, les placements à long terme permettent d'absorber la volatilité et augmentent les possibilité de rendement.

- **Les contrats multisupports**

Les contrats multisupports sont des contrats mixes pouvant contenir des supports à la fois en euros et en UC. Les primes versées sont reparties sur les deux fonds selon la convenance de l'assuré. Ces contrats représentent donc un bon équilibre entre le risque et le rendement et permettent de satisfaire différents profils d'épargnants.

Par la suite, nous n'étudierons que les supports en euros qui représentent le type de contrat de notre portfeuille d'étude.

1.3 Les caractéristiques d'un contrat d'épargne en euros

Dans notre portefeuille d'étude, les assurés souscrivent individuellement à un contrat d'épargne en euros. Ainsi, il versent des primes de façon ponctuelles à l'assureur et ces primes sont investies sur des fonds euros sécuritaires. L'assuré est soumis à des frais pour la gestion de son dossier et de son patrimoine. L'assureur promet de verser un taux minimum (positif ou nul) à l'assuré à chaque fin d'année, ainsi qu'une part des bénéfices réalisés à l'aide des placements de l'assuré. Ce type de contrat comprend donc les principales caractéristiques suivantes : les primes, les chargements et frais, le TMG, et la PB.

- **Les primes**

Les primes versées en assurance représentent l'un des éléments les plus importants de la souscription d'une assurance. En effet, les primes correspondent aux montants que l'assuré doit payer à l'assureur pour bénéficier de la couverture d'assurance. Le montant de la prime dépend de plusieurs facteurs tels que la nature de la couverture, le montant de la garantie, le niveau de risque, l'âge et la santé de l'assuré.
Le versement de la prime peut être effectué en une seule fois au début du contrat ou alors en plusieurs fois pendant la durée du contrat.
Il est important de mentionner que l'assureur ne peut en aucun cas exiger le paiement des primes (article L.132-20 du Code des Assurances).

- **Les frais**

Les frais représentent des sommes additionnels à la prime que l'assuré doit payer pour assurer l'entretien de son épargne et la gestion de son placement.
Les frais en assurance désignent les coûts liés à la souscription et à la gestion d'un contrat d'assurance. Ces frais peuvent inclure différentes composantes, telles que les frais d'adhésion à un contrat d'assurance, les frais de gestion de ce contrat, les frais de versement, les frais de sortie ou de rachat, les frais de transfert, les frais de courtage, les frais de garantie ou de cautionnement, etc.

- **Le Taux Minimum Garanti (TMG)**

Le taux minimum garanti ou taux technique est le taux minimum qui sera servi à la fin de l'année. Ce taux est fixé chaque année par l'assureur et doit être en accord avec la réglementation. L'assureur garantit donc au souscripteur un rendement annuel au moins égal au TMG. Dans le cas où le rendement des investissements effectués par l'assureur ne suffirait pas à couvrir le TMG, ce dernier devra puiser par ses ressources ou utiliser un autre moyen pour servir au moins le TMG à l'assuré. C'est pourquoi, le montant du TMG est également plafonné (articles A.132-1 et A.132-1-1 du Code des assurances) par la loi pour éviter des engagements excessifs de la part des assureurs.

- **La Participation aux bénéfices**

L'assureur garantit certes le versement d'un taux minimum fixé dans le contrat, mais a la possibilité de verser un taux plus élevé en cas de bons rendements. En effet, les assureurs s'engagent à redistribuer la majeure partie des bénéfices réalisés grâce aux primes des assurés. Il s'agit de la participation aux bénéfices. Chaque année, l'assureur a l'obligation de verser au moins 85% des bénéfices

financier et 90% des bénéfices techniques résultant de la différence entre les frais réels de la compagnie et ceux qu'elle a prélevés.

L'assureur a la possibilité de ne pas verser la PB en totalité et de la stocker dans une provision appelée PPB (Provision de Participation aux bénéfices) ou PPE (Provision pour participation aux Excédents). Les sommes mises en PPE doivent obligatoirement être distribuées aux assurés au bout de huit ans et permettent à l'assureur de lisser son taux servi. Ainsi, même lors d'une année avec un mauvais rendement, il pourra puiser des réserves dans la PPE pour servir un taux assez satisfaisant.

Le taux technique et la PB sont donc les deux éléments qui permettent la revalorisation du contrat. En outre, plusieurs risques sont inhérents à un contrat d'épargne individuelle, notamment les risques de rachat ou encore de décès. Ceux-ci seront présentés lors du chapitre sur la modélisation du passif.

Ainsi, après avoir présenté succinctement les caractéristiques du contrat d'épargne euros étudié, nous décrirons dans la suite de ce mémoire le modèle de projection utilisé.

2 Modélisation de l'actif

Afin de remplir son engagement envers l'assuré et de garantir un bon rendement, l'assureur se doit de mettre en place toutes les dispositions nécessaires pour protéger ses investissement, d'où la mise en place d'un modèle actif.
L'actif représente les différents placements effectués par l'assureur pour faire fructifier l'épargne. Dans le cas du portefeuille que nous étudions, les actifs sont composés en majorité d'actifs réels libellés en euros, essentiellement de type obligataires, actions et immobiliers.

Ces actifs doivent être au moins égaux à l'ensemble des engagements pris par l'assureur et sont soumis à des contraintes légales indiquées dans les articles R.332-1 et R.332-2 du Code des Assurances.

La modélisation de ces actifs repose sur un Générateur de Scénarios Économiques (GSE) servant à générer plusieurs environnements macro-économiques possibles appelées trajectoires. Le GSE permet de projeter les actions, l'immobilier, les taux réels ainsi que les taux nominaux. Les projections s'effectuent sur 50 ans avec un pas de temps annuel. Les éléments ressortis de ce GSE ont donc une importance capitale pour déterminer le rendement des actifs et donc le résultat de l'entreprise.

Compte tenu de l'intérêt du GSE dans nos travaux, nous décrirons son fonctionnement qui se base principalement sur la modélisation stochastique de certains facteurs de risques comme la courbe des taux ou encore l'indice de croissance des actions.
Mais avant cela, il est important de savoir que ce GSE repose sur un environnement de risque neutre et de comprendre de quoi il s'agit.

- **L'environnement risque neutre**

L'univers risque neutre est un univers virtuel dans lequel les primes des risques sont nulles et les espérances de rentabilité sont égales au taux sans risque. Les conditions permettant l'approche risque neutre sont l'Absence d'Opportunité d'Arbitrage (AOA) et la complétude de marché. Intuitivement, cela veut dire que dans cet univers, le rendement d'un actif en cas d'absence d'arbitrage doit être le

taux sans risque. De ce fait, la probabilité des événements futurs dans le monde risque neutre Q diffère de la probabilité historique $\mathbb{P}$ qu'on observe dans le monde réel.

Nous rappelons également que X_t, processus aléatoire dynamique est une martingale pour la filtration F_t (information connue à l'instant t) sous la loi de probabilité Q si : $\mathbb{E}_Q(X_{t+1}|F_t) = X_t$.
Ainsi, sous la probabilité risque neutre, tous les processus de prix évoluent en moyenne au taux sans risque.

Revenons à notre générateur de scénarios économiques. Celui-ci modélise les variables financières afin de reconstituer un environnement économique de marché dans lequel s'inscrira ensuite le modèle de gestion actif/passif. Nous allons maintenant présenter une méthode de modélisation des différentes classes d'actifs.

2.1 Modélisation des actions

Nous allons maintenant aborder la technique de modélisation utilisée pour les actions. Cette dernière repose sur un modèle de Black & Scholes dans un univers risque neutre. Le modèle de Black & Scholes est un modèle couramment utilisé en mathématiques financières et qui permet d'estimer en théorie la valeur d'une option de type européenne en fonction de différentes données de cette option et de l'actif sous-jacent sur lequel elle est adossée.

L'indice de croissance des actions est modélisé à l'aide d'un mouvement brownien géométrique modifié dans lequel le drift est égal au taux court généré à l'aide du modèle de taux nominaux que nous verrons dans la suite. Le cours des actions vérifie le processus suivant :

$$\frac{dS_t}{S_t} = r_t dt + \sigma W_t$$

Avec,

- S_t la valeur de l'action à l'instant t
- r_t l'espérance de rentabilité de l'action
- σ la volatilité de l'action
- W_t le mouvement brownien géométrique sous la probabilité historique $\mathbb{P}$

La solution explicite à cette équation est la suivante :

$$S_t = S_0 \exp\left(\left(r_t - \frac{\sigma^2}{2}\right) t + \sigma W_t\right)$$

De plus, la valeur boursière des actions est calculée comme suit :

$$VB_t = \left(1 + \frac{S_t}{S_{t-1}}\right) \times VB_{t-1}$$

2.2 Modélisation de l'immobilier

Tout comme pour les actions, le modèle utilisé pour l'immobilier est un modèle de Black & Scholes. L'indice immobilier suit la dynamique suivante :

$$\frac{dI_t}{I_t} = r_t dt + \sigma_t W_t$$

Avec,

— I_t la valeur de l'indice à l'instant t

— r_t l'espérance de rentabilité de l'indice

— σ la volatilité de l'indice

La calcul de la valeur boursière de l'immobilier se fait de la même façon que celui des actions, en remplaçant S_t par I_t dans la formule donnée plus haut.

2.3 Modélisation des taux

La modélisation des prix des zéro-coupon et obligations passe par l'utilisation des taux courts. Dans cette partie nous allons présenter la méthode utilisée pour modéliser les taux courts.

2.3.1 Taux d'intérêt nominaux

La modélisation des taux d'intérêt nominaux repose sur un modèle LMM (Libor Market Model). Il permet de reconstruire parfaitement la courbe des taux et de reproduire les prix d'options sur taux. Cette méthode suppose que les taux forward suivent le processus stochastique suivant :

$$\frac{dF_k(t)}{F_k(t)+\delta} = V(t) \sum_{i=m(t)}^{k} \left[\frac{\Delta(F_i(t)+\delta)}{1+\Delta(F_i(t)} \gamma_i(t).\gamma_k(t) \right] dt + \sqrt{V(t)}\ \gamma_k(t) dZ^d(t)$$

Avec,

— $F_k(t)$ le taux forward en t sur la période $(T_k; T_{k-1})$

— $\gamma_i(t)$ la composante de volatilité du taux forward qui dépend de la durée jusqu'à l'échéance

— $V(t)$ la variance stochastique

— $Z^d(t)$ le mouvement brownien géométrique

— δ le coefficient de déplacement

— $m(t)$ le plus petit entier tel que $t \leq m(t)$

2.3.2 Taux d'intérêt réels

Les taux d'intérêts réels sont modélisés via un modèle de Vasicek à deux facteurs. Ce modèle se présente comme la somme d'un premier facteur d_r représentant le taux d'intérêt court instantané et d'un second facteur d_m représentant le taux d'intérêt à long terme. C'est un modèle qui permet de reconstituer l'ensemble de la courbe des taux par formule fermée au moyen de la valeur du taux court. La dynamque du taux court de ce modèle est la suivante :

$$dr(t) = a_1\ (m(t) - r(t))dt + \sigma_1 W_1(t)$$
$$dm(t) = a_2\ (-m(t))dt + \sigma_2 W_2(t)$$

Avec,

— $r(t)$ le taux d'intérêt court terme

— $m(t)$ le taux d'intérêt long terme

— la moyenne de long terme

— a_1 et a_2 les vitesses de retour à la moyenne

— σ_1 et σ_2 les volatilités instantanées

— $W_1(t)$ et $W_2(t)$ des mouvements browniens géométriques indépendants

2.3.3 L'inflation

L'inflation est obtenue à l'aide des taux nominaux et réels comme l'indique la formule suivante :

$$Inflation_t = \frac{P_r(t-1,t)}{P_n(t-1,t)} - 1$$

Avec,

— $P_r(t-1,t)$ le prix d'une obligation zéro-coupon nominale de maturité un an à la date $t-1$

— $P_n(t-1,t)$ le prix d'une obligation zéro-coupon réelle de maturité un an à la date $t-1$

2.4 Modélisation des instruments de taux

Dans cette partie nous aborderons la modélisation des obligations à taux fixe et des zéro-coupon.

- **Flux de coupon**

Le flux F_i d'un taux fixe ou zéro-coupon à la date T_i et pour une périodicité annuelle est :

$$F_i = C, \quad \forall i = 1, ..., N-1$$
$$F_N = 1 + C$$

Où C est le taux de coupon risque neutre.

- **Flux de la période**

Le flux en t est calculé à l'aide de la somme du coupon et de la valeur de remboursement en t :

$$Flux_t = Coupon_t + Remboursement_t$$

Sachant que :

$$Coupon_t = C \times VRemb_t$$
$$Remboursement_t = VRemb_t$$

Où $VRemb_t$ est la valeur de remboursement en t.

- **Valeur bilan**

La valeur bilan d'une obligation à taux fixe est calculée de la façon suivante :

$$VBilan_t = \sum_{i=1}^{N} \frac{F_i}{(1+r)^{T_i}} \times VRemb_t$$

Où r est le taux de rendement actuariel.

- **Valeur de marché**

La valeur de marché des instruments correspond à la somme des flux projetés (coupons et remboursements à échéance) jusqu'à maturité, ces derniers étant au préalable actualisés avec la courbe zéro-coupons. La formule de la valeur de marché est la suivante :

$$VMarché_t = \sum_{i=1}^{N} \frac{F_i}{(1 + R(t, T_i))^{T_i}} \times VRemb_t - CC_t$$

Avec,

— $R(t, T_i)$ le taux zéro-coupon de maturité T_i observé en date t

— N la maturité

— CC_t le coupon couru en t

2.5 Stratégie de revalorisation

La stratégie de revalorisation consiste à déterminer la production financière réalisée par l'actif à la fin de chaque année de projection, et à la répartir entre la participation aux bénéfices et les chargements.

Notons que la stratégie de revalorisation des contrats mise en oeuvre au sein du modèle de projection sur le périmètre épargne euros comprend :

— la satisfaction des garanties de taux

— le versement de la participation aux bénéfices discrétionnaire

— la satisfaction des clauses de participation aux bénéfices contractuelles

— la satisfaction de la contrainte du minimum de participation aux bénéfices réglementaire

La stratégie de revalorisation appliquée à l'ensemble des trajectoires de la simulation stochastique s'effectue en deux temps. Tout d'abord, sachant que les prestations telles que les décès et les rachats, sont programmées pour se réaliser en milieu d'année, la part de revalorisation sur la demi-période qui est servie avec la prestation est donc un pourcentage du taux servi pour le stock de contrats au cours de l'exercice précédent.

En fin d'année, le modèle calcule les produits financiers totaux du portefeuille. Ces produits financiers seront diminués des intérêts crédités au taux technique et de la participation aux bénéfices. Pour finir, la charge de participation aux bénéfices au-delà du taux technique est calculée avec une cible de revalorisation globale dépendant notamment du taux moyen d'emprunt d'État (TME) actuel et passé, en tenant compte des chargements sur encours et produits financiers prévus au contrat.

Dans un second temps, une comparaison est effectuée entre d'une part les produits financiers totaux nets et la cible de revalorisation. La PPE peut être utilisée dans le cas où les produits financiers ne permettent pas d'assurer le paiement des prestations et la revalorisation des encours. Des richesses latentes peuvent également être réalisées au besoin. De plus, dans le cas où les ressources disponibles demeurent insuffisantes et afin de satisfaire une revalorisation minimale, des coûts sur fonds propres

peuvent aussi être générés.

Notons que le financement de la partie non-discrétionnaire des contrats (contraintes obligatoires) correspond aux éléments qui doivent obligatoirement être financés et qui peuvent donc mener à la réalisation de coût fonds propres. La PPE ne peut pas être utilisée pour réaliser des contraintes obligatoires.

Ainsi, après s'être confronté à la modélisation des différents éléments de l'actif, nous allons nous attarder sur la modélisation des éléments du passif.

3 Modélisation du passif

Dans cette partie, nous allons étudier comment sont modélisés et projetés les différents éléments du passif. Nous aborderons notamment les prestations (décès et rachats), les frais, les différentes provisions et le BE. Nous rappelons que chaque élément du bilan sur un horizon de 50 ans avec un pas de temps annuel.r

3.1 Les prestations

Afin de pouvoir évaluer les engagements de l'assureur, il est impératif de pouvoir projeter les flux des prestations composés notamment des décès et des rachats, sur l'horizon du contrat. Ces flux dépendent de l'évolution des caractéristiques des contrats et des assurés au cours du temps. Notons que les prestations sont des flux de sorties en milieu d'année. Elles sont accompagnées d'une diminution des provisions mathématiques après revalorisation au titre des intérêts crédités et des taux garantis.

3.1.1 Décès

Le décès de l'assuré avant l'échéance est considéré comme une sortie du contrat. Dans ce cas, l'assureur va verser au bénéficiaire un capital égal au montant acquis au jour du décès.
Les sorties pour cause de décès sont calculées comme un taux de mortalité dépendant de l'âge de l'assuré, appliqué à la PM de demi-année.

Les taux de mortalité sont estimés à partir de table de mortalité d'expérience calibrées sur les données historiques propres au périmètre épargne.
Pour un âge donné x, cette table de mortalité donne le nombre de personnes en vie l_x d'âge x . On peut alors en déduire le taux de mortalité q_x d'un individu d'âge x :

$$q_x = 1 - \frac{l_x + 1}{l_x}$$

On peut alors calculer le montant des décès à l'aide de ce taux et du niveau de provisions mathématiques :

$$Prestation_{décès} = q_x \times PM$$

3.1.2 Rachats

Un contrat d'épargne en euros est souscrit pour une durée plus ou moins longue, généralement au moins huit ans pour pouvoir bénéficier de la fiscalité avantageuse associée à ce type de contrat. Les souscripteurs ont tout de même la possibilité de procéder au rachat de leur contrat avant l'échéance de celui-ci. Ce rachat leur permet de récupérer totalement ou partiellement leur part de provision mathématique (valeur de rachat).
On peut donc distinguer deux types de rachats :

— le rachat partiel qui consiste à récupérer auprès de l'assureur une partie de son épargne avant l'échéance du contrat

— le rachat total qui permet de récupérer la totalité de son épargne et met ainsi fin au contrat.

Par ailleurs, l'assuré peut également demander une avance auprès de l'assureur, qui correspond à un prêt qu'il devra rembourser par la suite moyennant intérêts.

Dans le modèle, les deux types de rachat sont pris en compte. C'est pourquoi le montant de rachat correspond à la somme des taux de rachats partiels (supposés uniquement structurels) et totaux (structurels et dynamiques) appliquée à la PM après le calcul des décès.

Ainsi, en plus des rachats structurels que l'assureur peut observer dans un contexte économique normal sur les contrats d'assurance vie épargne euros, l'assureur doit tenir compte de rachats conjoncturels. Ce type de rachat intervient dans un contexte fortement concurrentiel lorsque l'assuré arbitre son contrat au profit d'autres supports financiers.

- **Rachat structurel**

Le rachat structurel (partiel et total) est lié au fait que les assurés puissent avoir besoin de liquidités de manière générale. Ces rachats peuvent s'expliquer en grande partie par l'ancienneté des contrats. En effet, au-delà de huit ans, les rachats augmentent fortement du fait de la défiscalisation partielle à partir de cette date.

- **Rachat conjoncturel**

Le rachat conjoncturel représente la part de rachat induite par le comportement des assurés, en réponse à l'écart constaté entre le taux servi par leur assureur et les taux offerts sur le marché par la concurrence.
La modélisation des rachats dynamiques (conjoncturels) repose sur la notion de taux attendu par les assurés en terme de revalorisation. Si le taux attendu est supérieur au taux servi, le risque que les assurés rachètent leur contrat sera plus important l'année suivante. À l'inverse, si le taux servi est supérieur au taux attendu, les assurés auront tendance à moins racheter. Ce phénomène traduit un réinvestissement de la part des assurés sur les produits dont la performance est supérieure à leurs attentes. De plus, en fonction des catégories d'assurés, l'attente en termes de rendement par rapport au taux de marché sera supposé plus ou moins haute.

3.2 Frais et Chargements

Le montant des frais pour chaque exercice est projeté en tenant compte de la destination des dépenses (acquisition, gestion, administration). Les chargements permettent de compenser les frais et sont projetés de la même façon, suivant 3 catégories :

— Les chargements sur primes qui proviennent des conditions contractuelles et correspondent aux prélèvements effectués pour couvrir éventuellement les frais d'acquisition et les commissions.

— Les chargements sur primes qui sont calculés en pourcentage des primes de la période.

— Les chargements sur encours ou sur production financière qui correspondent aux prélèvements effectués pour gérer les contrats.

Les chargements sur primes (souscriptions, versements libres, versements programmés) sont calculés en pourcentage des primes de la période. Les chargements sur encours ou sur production financière correspondent aux prélèvement effectués pour gérer les contrats.

3.3 Les provisions techniques

3.3.1 La Provision mathématique

La provision mathématique (PM) se définit comme la différence entre la valeur actuelle probable des engagements de l'assureur et la valeur actuelle probable des engagements de l'assuré :

$$PM = VAP(assureur) - VAP(assuré)$$

Dans le cadre des contrats d'épargne en euros, la PM représente l'épargne acquise revalorisée des assurés. Cette provision est évaluée à chaque date de clôture en étant revalorisée au taux servi cible de l'année. Elle correspond à l'ensemble des flux entrant diminué de l'ensemble des flux sortant. Nous avons alors :

$$PM_{cloture} = (PM_{ouverture} - décès - rachats - chargements) \times taux_{revalorisation}$$

La PM est la provision la plus connue mais nous trouvons d'autres provisions techniques modélisées au sein du modèle de projection.

3.3.2 La provision pour paticipation aux excédents

La provision pour participation aux excédents (PPE) est comme nous l'avons défini précédemment, une réserve des bénéfices réalisés par l'assureur et non redistribués immédiatement aux assurés.
Cette provision permet à l'assureur de lisser dans le temps la rémunération de ses contrats et ainsi de servir un taux régulier auprès des assurés. Par exemple, les années où les produits financiers seront importants, l'assureur pourra servir son taux cible et mettre le surplus dans la réserve de PPE.
Cette dotation à la PPE pourra être utilisée les années suivantes (dans la limite de huit ans) lorsque les produits financiers seront moins importants et insuffisants.
Les stratégies financières mises en place dans le but de gérer la PPE (montant à doter en PPE et montant à reprendre) constituent un point important pour l'assureur.

3.4 Modélisation du Best Estimate

3.4.1 Principe

La meilleure estimation de l'engagement de l'assureur envers ses assurés correspond à la moyenne pondérée par leur probabilité des flux de trésorerie futurs, compte tenu de la valeur temporelle de l'argent. Cette partie du passif est appelée best estimate liabilities ou plus simplement best estimate. Le BE est basé sur la projection des cash-flows futurs d'actif et de passif. Il est égal à la valeur actuelle probable des flux de trésorerie générés par le contrat. De plus, l'estimation des flux futurs de trésorerie est effectuée à partir de données fiables, réalistes et cohérentes avec le marché.

Afin de capter au mieux la valeur temps des options et garanties du contrat, une évaluation du BE selon une modélisation stochastique est requise. Il a été choisi de simuler 1 000 scénarios de marché pour réaliser les 1 000 projections. Notons que pour chaque scénario, les étapes suivantes sont mises en place :

— Modélisation stochastique de l'actif dans un univers risque neutre

— Modélisation stochastique du passif dans un univers risque neutre

— Modélisation des interactions actif/passif

— Projection des cash-flows sur la durée de projection

— Somme et actualisation des cash-flows pour toutes les dates de projection

Nous obtenons alors un BE par scénario en actualisant les flux de trésorerie avec la courbe des taux sans risque correspondante. Pour obtenir le BE final, il suffit alors de faire la moyenne des BE provenant de chaque scénario :

$$BE = \frac{1}{S} \sum_{k=1}^{S} \sum_{i=1}^{N} \frac{CF_i^k}{(1+r_i)^i}$$

Avec,

— S le nombre de scénarios (ici 1 000)

— N la date d'extinction du portefeuille (la projection des flux attendus est réalisée en run-off). Dans notre cas, N est fixé à 50 ans.

— r_i le taux d'actualisation de la courbe des taux retenue correspondant à un engagement de durée i

— CF_i^k le cash-flow de l'année i pour le scénario k

Comme énoncé ci-dessus, la valorisation du BE impose de projeter les flux futurs de trésorerie. C'est pourquoi, nous devons déterminer les lois de comportement des assurés et de l'assureur ainsi que le taux d'actualisation utilisé.
Le comportement de l'assuré comprend les décès, les rachats ou encore les versements libres tandis que le comportement de l'assureur se compose seulement de sa politique de participation.

Notons que la loi de mortalité des assurés est déterministe. Pour cela, nous avons retenu une certaine table de mortalité.

3.4.2 Modélisation des rachats

Comme expliqué dans la section sur les prestations, deux types de rachats sont modélisés : les rachats structurels et les rachats conjoncturels. Cette modélisation est faite telle que le total des prestations de rachat soit calculé comme :

$$Rachat\ total = Rachat\ structurel + Rachat\ conjoncturel$$

- **Rachat structurel**

La construction des lois de rachat structurel se base sur l'historique des assurés ayant racheté la totalité de leur contrat.

- **Rachat conjoncturel**

Comme indiqué précédemment, la modélisation des rachats dynamiques repose sur la notion de taux attendu par les assurés en terme de revalorisation.
Les rachats conjoncturels sont couramment modélisés par une fonction dépendant uniquement de l'écart entre le taux servi et le taux dépendant de l'environnement économique.

L'ACPR propose dans le QIS 5, une fonction pour modéliser les rachats conjoncturels. Le taux attendu proposé par l'ACPR pour modéliser les rachats conjoncturels étant le TME, nous avons :

$$Rachats\ conjucturels = \begin{cases} RC_{max}\ si\ R - TME < \alpha \\ RC_{max}\frac{R-TME-\beta}{\alpha-\beta}\ si\ \alpha < R - TME < \beta \\ 0\ si\ \beta < R - TME < \gamma \\ RC_{min}\frac{R-TME-\gamma}{\delta-\gamma}\ si\ \gamma < R - TME < \delta \\ RC_{min}\ si\ R - TME < \delta \end{cases}$$

Avec,

— R le taux global de rendement (TGR)

— TME le taux moyen d'emprunt d'Etat

— RC_{max} le taux de rachats dynamiques maximum

— RC_{min} le taux de rachats dynamiques minimum

— α le seuil en-dessous duquel la loi de rachat dynamique est constante et égale à RC_{max}

— β et γ les bornes inférieure et supérieure d'un intervalle où le comportement des assurés est indépendant de l'écart entre R et le TME

— δ le seuil au-delà duquel les rachats dynamiques sont constants et égaux à RC_{min}

L'ACPR a par ailleurs fixé des valeurs minimales et maximales pour l'ensemble de ces paramètres, définissant ainsi une loi « plancher » et une loi « plafond » permettant de modéliser les rachats conjoncturels :

α	β	γ	δ	RC_{max}	RC_{min}
-5%	-1%	1%	3%	30%	-5%

Notons aussi qu'il existe une autre méthode de calcul de rachat conjoncturel qui consiste à appliquer un coefficient multiplicateur au taux de rachat structurel. Ce coefficient va dépendre de l'écart entre le TGR servi l'année précédente et la moyenne des derniers TME.

Cette loi de rachat dynamique est donnée par :

$$Taux\ rachats\ dynamiques = min(taux_{max}, max(0, Ecart))$$

Avec,

— $Ecart = fonction(b, c, TME, t, n)$

— $taux_{max}$ le taux de rachat maximum (en %)

— a un coefficient

— b le seuil de déclenchement (entre 0% et 5%)

— c la part du TME retard pour l'écart (en %)

— n le nombre d'années d'historiques du TME

3.4.3 La politique de participation aux bénéfices

Afin de valoriser au mieux l'épargne du client, il est prévu un taux global de revalorisation cible. Nous obtenons alors la PB cible que l'assureur souhaite distribuer à ses assurés :

$$PB_{cible} = TGR_{cible} \times PM_{1/2}$$

Avec,

— $PM_{1/2}$ la provision mathématique de demi-année.

4 Interaction entre l'actif et le passif

Le modèle Actif/Passif, utilisé dans cette étude, permet de relier la dynamique du passif à celle de l'actif et peut être utilisé selon une approche déterministe ou stochastique où les flux de passif et d'actif sont projetés simultanément. Dans le cas de projection stochastique le modèle actif est soumis à plusieurs scénarios économiques. Les interactions Actif/Passif concernent notamment la revalorisation des contrats, les rachats dynamiques et la stratégie financière.

Notre modèle de projection distingue trois étapes de calcul à chaque pas de projection.

1. En début d'année, les portefeuilles d'actifs et de passifs sont mis à jour

2. En milieu d'année, plusieurs niveaux de calcul sont mis en place :

— au passif, une revalorisation des contrats au niveau des taux minimums garantis sur la première demi-année (y compris les intérêts crédités) est faite. A ce même niveau, le paiement des prestations, la collecte des primes (nouveaux contrats, versements libres et programmés pour les contrats existants), des chargements sur primes ainsi que le paiement des commissions sur primes sont effectués ;

— à l'actif, nous observons la tombée des flux d'actifs tels que les dividendes, les coupons ou les remboursements

— au niveau des interactions actif/passif, la revalorisation discrétionnaire des sorties est faite. Le flux global correspondant à la somme des flux d'actifs et de passifs perçus et payés est investi ou désinvesti en fonction de la stratégie financière

3. En fin d'année, il y a également plusieurs niveaux de calcul :

— au passif, nous calculons les différentes masses à financer relatives aux objectifs de revalorisation des contrats. La revalorisation des contrats au niveau des taux minimums garantis sur la deuxième demi-année (y compris intérêts crédités) est mise en place

— à l'actif, nous calculons les masses financières disponibles dans le cadre de la stratégie de revalorisation discrétionnaire (plus-values, produits financiers)

— au niveau des interactions actif/passif, la stratégie de participation aux bénéfices discrétionnaire est mise en place. De plus, il faut vérifier que les garanties contractuelles de revalorisation des contrats ont bien été respectées et réaliser les retraitements le cas échéant, mais aussi vérifier du respect de la contrainte réglementaire du minimum de participation aux bénéfices et exécuter la stratégie financière de fin d'année

Ci-dessous un schéma récapitulatif de ce modèle d'interaction actif/passif :

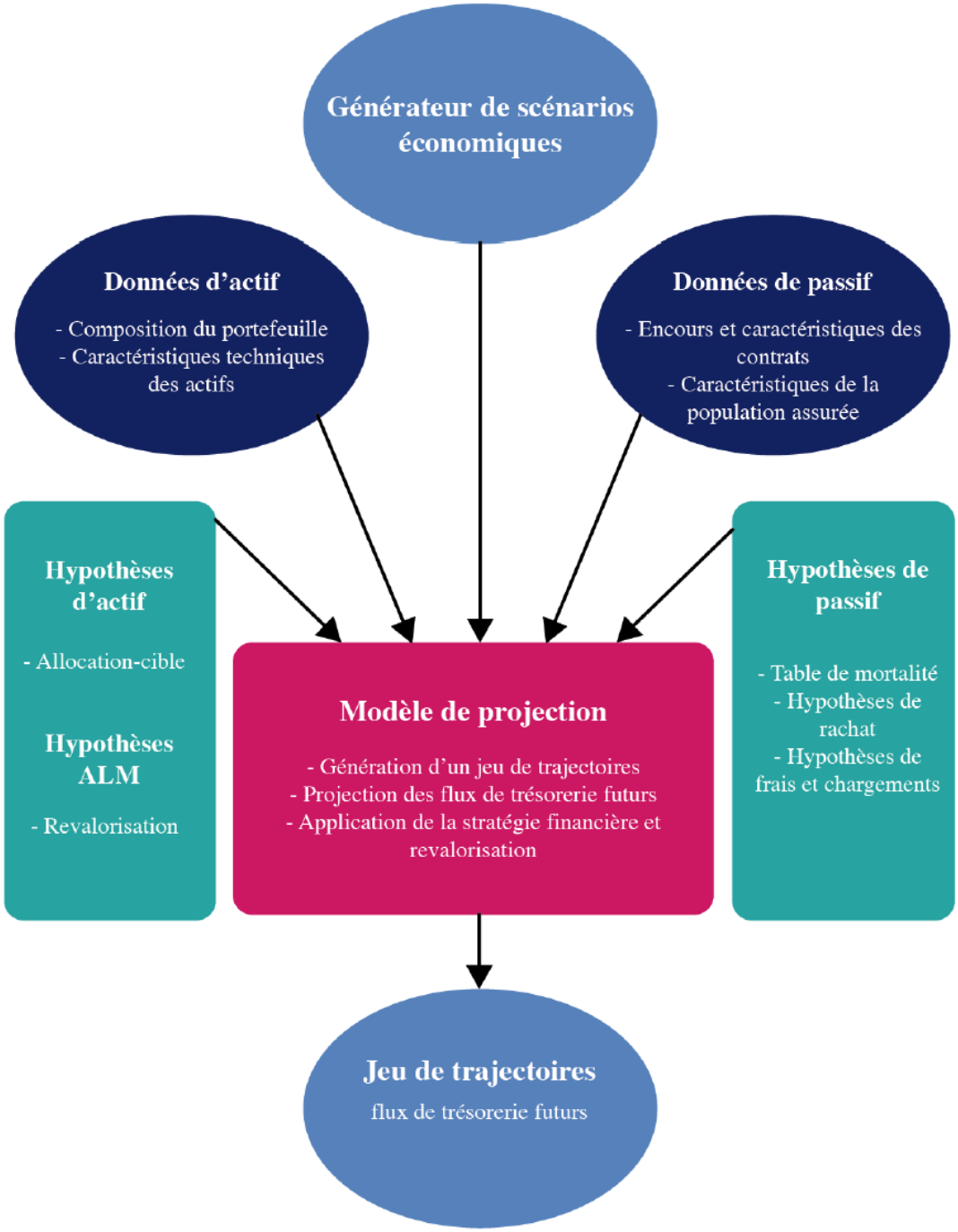

FIGURE 3.1 – Illustration du modèle de projection actif/passif

Chapitre 4

Conception de la base d'apprentissage

1 Méthodologie de construction de la base

A présent que nous savons sur quoi est basé un modèle d'épargne, nous détenons toutes les clés pour le choix des variables utiliser. Dans cette partie nous présenterons les données utilisées, et les variables que nous avons choisies pour notre modèle.

Dans cette partie, nous présenterons la façon dont les données ont été collectées et la base de données construites. Nous parlerons principalement des observations qui constituent notre base. Il est donc essentiel de faire la différence entre les variables et observations.
Une observation représente un individu la population étudiée tandis qu'une variable représente les caractéristiques de cet individu. Il s'agit donc des lignes de la base de données.

Lors d'un exercice de production, l'assureur est emmené à effectuer des simulations pour le calcul de ses résultats à partir d'un modèle complexe implémenté dans l'application NEMO (Application de modélisation interne de CNP Assurances). Ce modèle ressort en résultats de nombreuses variables parmi lesquelles la VIF. Notre objectif étant de reproduire les méthodes de calcul de ce modèle à l'aide du Machine Learning, nous sommes servis des différentes simulations effectuées sur NEMO.

Pour cela, nous avons effectué plusieurs simulations puis récupéré les différents paramètres utilises en entrée du modèle et les VIF ainsi que ses sous variables ressorties en résultats du modèle. Les variables d'entrée du modèle nous servirons à créer l'ensemble de nos variables explicatives et les VIF et sous variables à créer les variables à expliquer.
La complexité de cet exercice réside dans le fait d'agréger les différentes données d'entrées qui se présentent toutes sous un format différent en un seul tableau de la même structure.

L'application NEMO permet de faire des modélisations par approche stochastique. Pour produire un résultat donné, le modèle effectuera la moyenne des résultats de 1000 scénarios stochastiques qui sont en réalité dérivés du GSE. Nous avons donc fait le choix de récupérer non pas les informations du résultat moyen mais celui des 1000 scénarios.

De plus, pour chaque simulation donnée, nous avons la possibilité d'observer non pas seulement le résultat Central mais aussi le résultat de l'environnement suite aux différents chocs S2. Nous avons donc fait le choix d'enrichir notre base en prenant pour chaque simulation les 11 chocs S2 suivants :

- Central
- Choc Action Type 1
- Choc Action Type2
- Choc Hausse Taux
- Choc Baisse Taux
- Choc Immobilier
- Choc Mortalité
- Choc Hausse Rachats
- Choc Baisse Rachats

— Choc Longévité

— Choc Frais

— Choc Inflation

D'autre part, nous avons effectué plusieurs jeux de sensibilités touchant à la fois les hypothèses économiques, techniques et financières. Ceci dans le but de rajouter à notre base suffisamment d'information pour capter la sensibilité du modèle Actif/passif dans différents environnements et configurations.

Notre principal axe de simulation est le Closing, représentant les résultats en fin d'année. Ensuite, nous avons appliqué au Closing deux sensibilités sur les taux :

— Closing_Tx_M50bps : Environnement Closing avec une variation de -50bps sur les taux

— Closing_Tx_P50bps : Environnement Closing avec une variation de +50bps sur les taux

A ce moment-là nous disposions alors de 3x11x1000 = 33000 observations. Nous avons effectué des tests sur cette base mais il s'est avéré qu'elle ne contenait pas assez d'informations pour nous fournir des résultats satisfaisants.
Nous avons donc décidé de l'enrichir en effectuant de nouvelles sensibilités croisées à la fois sur le Closing et sur les en environnements Closing_Tx_M50bps et Closing_Tx_P50bps. Pour chaque environnement, nous avons effectué 10 sensibilités, ce qui nous fait un total de 11 simulations par environnement.

Le tableau ci-dessous résume l'ensemble des sensibilités appliquées et des simulations effectuées sur l'environnement Closing :

Environnement	Simulation effectuée
Closing - Valeur Centrale de l'année N après mise à jour de l'environnement économique	scénario de référence
	Sensibilités Hausse & Baisse des Coûts de +/- 5%
	Sensibilités Hausse & Baisse de la PPE de +/- 10%
	Sensibilités Hausse & Baisse des Rachats de +/- 10%
	Sensibilités Hausse & Baisse des Commissions de +/- 10%
	Sensibilité Hausse & Baisse des Chargements de +/- 10%

TABLE 4.1 – Ensemble des simulations et sensibilités

Les mêmes simulations ont été effectuées sur les deux autres environnements Closing_Tx_M50bps et Closing_Tx_P50bps. Nous avons donc un total de 33 simulations. Notre base finale compte 3 environnements x 11 sensibilités x 11 chocs S2 x 1000 scénarios, soit 363000 lignes.

2 Les variables explicatives

Nous rappelons que les variables explicatives sont les variables qui serviront à prédire la ou les variables cibles. Elles doivent donc communiquer suffisamment d'informations sur l'activité de l'assureur afin de pouvoir en déduire le résultat. Nous les avons regroupé en trois catégories : les variables contractuelles et du passif, les variables du GSE et les variables de l'actif.

2.1 Les variables contractuelles et du passif

D'une part, nous avons retenus une sélection des principales variables contractuelles représentants les caractéristiques du produit en question (hypothèse de chargement, de coûts, commissions et de TMG). Nous avons d'autre part identifié d'autres variables décrivant les hypothèses barométriques comme les hypothèses de rachats et de mortalité reflétant le comportement des assurés.
Une prise en compte du niveau de richesse initiale en particulier an lien avec la PPE a également été retenu.

Ces variables permettent ainsi de comprendre les différents facteurs qui pourraient impacter les contrats du portefeuille dans le temps.
Les variables en question sont les suivantes :

Code Variable	Description
Tx_CHGT	Taux de chargement : part de l'encours prélevée chaque année au titre de la rémunération de l'assureur
Tx_COMM	Taux de commissions : part de l'encours retenue chaque année au titre des commissions de l'assureur
Tx_COUT	Taux de Couts
Tx_RCHT	Taux de Rachat total
Tx_PPE	Niveau de PPE à l'encours
Tx_Décès	Taux de Décès
TMG	Taux Minimum Garanti des contrats

TABLE 4.2 – Tableau des variables contractuelles et du passif

Ci-dessous un aperçu de la répartition de quelques variables via un box-plot :

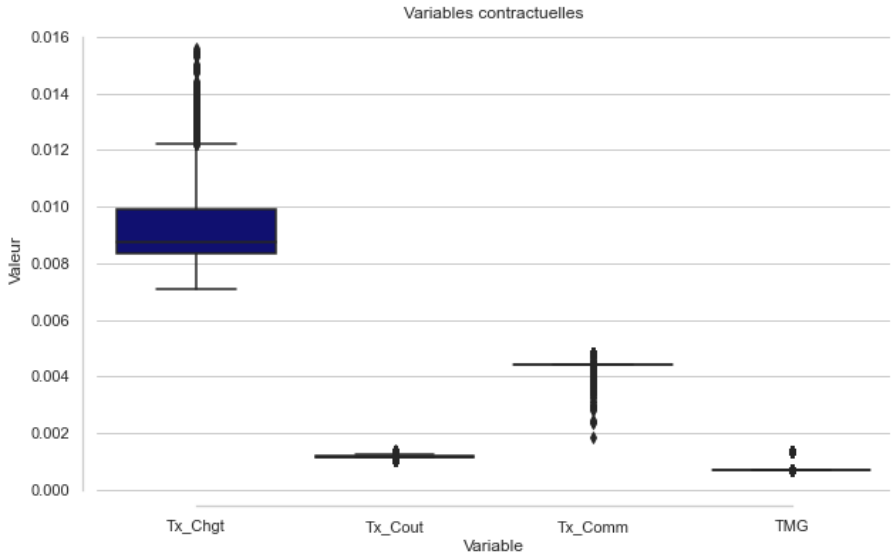

FIGURE 4.1 – Boxplot des variables contractuelles

Le Box-Plot ou boîte à moustache est un graphique permettant de visualiser la dispersion des variables en présentant la médiane, les quartiles (bords du rectangles) et les extrémités.

On constate une faible variation dans les variables taux de coût, commissions et TMG. En effet cela est dû au fait que nous travaillons sur un même portefeuille et donc les valeurs des variables contractuelles sont très similaires. Cette observation souligne l'utilité d'avoir effectué des sensibilités sur ces valeurs pour obtenir plus de variations.

Ce constat fera l'objet de travaux complémentaires dans un second temps afin d'enrichir la base d'apprentissage en lien avec les variables en question moyennant la simulation de plusieurs jeux de sensibilités.

2.2 Les variables du GSE

Les variables du GSE représentent les différentes valeurs en sorties du GSE et qui sont utilisées en input dans le cadre de l'exercice de modélisation classique de l'assureur. Il s'agit principalement de variables économiques, représentant les diverses possibilités d'évolution du marchés financiers dans les 50 années à venir. Elle sont importantes pour notre modèle car le rendement d'un produit épargne dépend en grande partie de la situation du marché des actions et des obligation dans lesquels l'assureur a investi les primes des assurés.

Le GSE permet de ressortir une multitude de variables mais ne pouvant pas toutes les prendre, nous avons identifié par avis d'expert, six indicateurs clé pour nos travaux :

— L'indice CAC
— L'indice Immobilier (Immo)
— Le Taux Zero Coupon 1 an (ZC_1Y)
— Le Taux Zero Coupon 10 ans (ZC_10Y)
— Le TME
— L'inflation (Inf)

Comme nous l'avons mentionné dans le chapitre 3, le GSE effectue des projections stochastiques pour 1000 scénarios et sur 50 ans. Chaque indicateur du GSE se présente donc sous la forme d'un tableau de 1000 lignes et 50 colonnes. Ci-dessous l'exemple du cas du TME :

Scénario	TME_1	TME_2	...	...	TME_50
1	- 0.59%	-0.52%	...	...	1.5%
2	- 0.56%	-1.16%	...	...	1.84%
...	...	...	...	...	...
1000	-0.64%	-0.50%	...	...	5.1%

TABLE 4.3 – Sorties TME du GSE

Où TME_x représente la projection du TME dans x ans.

Nous disposons donc de 50 variables pour chacun de ces indicateurs, soit 300 variables au total. Notons qu'un trop grand nombre de variables rendrait les temps de calculs beaucoup trop longs et pourraient empiéter sur la qualité du modèle en cas d'informations superflues.
D'un autre côté, nous ne pouvons pas nous permettre de négliger les informations qu'elles apportent. Nous avons donc opté pour deux mesures d'échantillon bien connues en statistique : la moyenne et la variance.

Afin de tirer le maximum d'informations et de limiter le nombre de variables du modèle, nous avons agrégé les variables par la moyenne et la variance comme présenté dans le tableau ci-dessous pour chaque indicateur.

Code variable	Description
Indicateur_Mean1Y5	Moyenne des projection de l'année 1 à 5 de l'Indicateur
Indicateur_Mean6Y10	Moyenne des projection de l'année 6 à 10 de l'Indicateur
Indicateur_Mean11Y20	Moyenne des projection de l'année 11 à 20 de l'Indicateur
Indicateur_Mean21Y30	Moyenne des projection de l'année 21 à 30 de l'Indicateur
Indicateur_Mean31Y50	Moyenne des projection de l'année 31 à 50 de l'Indicateur
Indicateur_Mean1Y50	Moyenne des projection de l'année 1 à 50 de l'Indicateur (toute la durée de projection)
Indicateur_Std1Y20	Ecart type des projection de l'année 1 à 20 de l'Indicateur
Indicateur_Std21Y50	Ecart type des projection de l'année 21 à 50 de l'Indicateur
Indicateur_Std1Y50	Ecart type des projection de l'année 1 à 50 de l'Indicateur

TABLE 4.4 – Agrégation des variables du GSE

Nous disposons donc de 9x6 = 54 variables économiques. Ci-dessous un graphique à titre indicatif de la distribution de nos variables pour le TME :

Ces figures nous montrent que le TME est une variable qui croit au court du temps. On constate

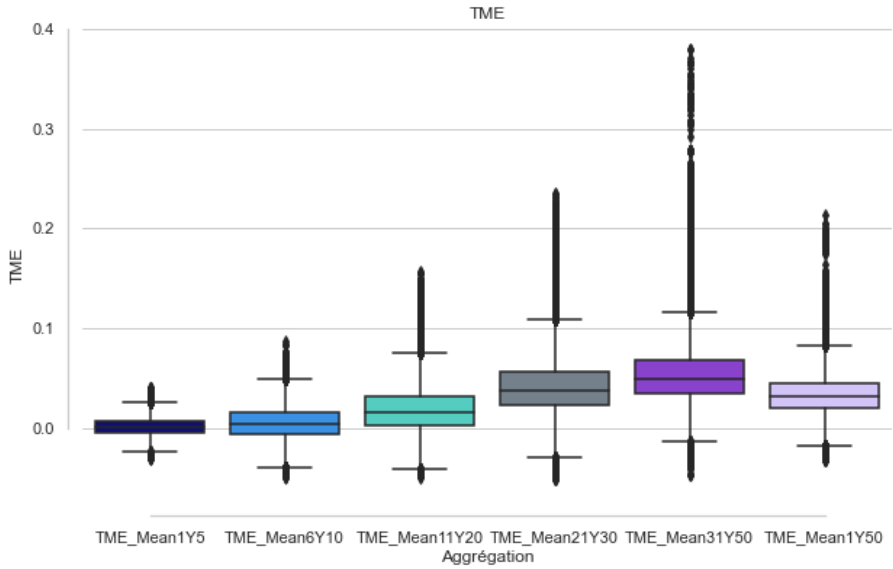

FIGURE 4.2 – Agrégation du TME par la moyenne

que la variable TME est relativement bien dispersée compte tenu la multitude des scénarios générés par notre GSE.

2.3 Les variables sur le portefeuille d'actif

Ces variables sont relatives à la stratégie d'investissement de l'assureur. Elles permettent de comprendre la façon dont le portefeuille d'actif a été reparti , que ce soit des actions, des obligations, de l'immobilier, etc. . .. Nous avons jugé intéressement d'inclure cet aspect dans notre modèle afin qu'il puisse également servir plus tard comme testeur lors de l'implémentation d'une nouvelle stratégie d'actif. Toutes fois, il faudrait pour cela bien étudier tous les facteurs clé et les paramètres nécessaires pour comprendre la stratégie d'investissement d'un assureur et les différents éléments qui l'impactent.

La principale donnée d'actif utilisées en input du modèle d'entreprise est un fichier Excel qui recense tous les actifs du portefeuilles détenus ainsi que leurs différents paramètres. Ainsi en ligne, on détient les actifs et en colonne les paramètres relatif à chaque actif.

Afin d'inclure cette donnée dans le même format que notre base de données, nous avons décidé d'agréger les actifs selon certains critères et d'observer trois indicateurs : la valeur boursière, les plus ou moins-value latentes et la duration.

2.3.1 Les indicateurs

Nous avons sélectionné trois indicateurs clé pour un actif :

- **La duration**

La duration est une mesure de la sensibilité des obligations aux variations des taux d'intérêt. Elle est calculée en tenant compte du taux d'intérêt, de la durée de vie restante de l'obligation et des flux de trésorerie qu'elle génère. Plus la duration est longue, plus l'obligation est sensible aux variations des taux d'intérêt. Si les taux d'intérêt augmentent, la valeur de l'obligation diminue et inversement. L'importance de la duration est qu'elle permet de mesurer le risque de taux d'intérêt associé à une

obligation. Une obligation avec une duration courte est généralement moins risquée qu'une obligation avec une duration longue.

- **La valeur boursière**

La valeur boursière (VBOURS) ou valeur de marché(VM) d'un actif, c'est à dire la valeur actuelle de l'actif en question. Elle s'oppose à la VNC (Valeur nette comptable) de l'actif qui représente le coût auquel l'actif a été acquis. En effet, la valeur d'un actif est emmené à évoluer au cours et différer de la valeur à laquelle il a été acquis.

- **Les PMVL**

Les PMVL (plus ou moins-values latentes) sont la différence entre le prix actuel d'un actif et son coût d'acquisition. On observe la relation suivante :

$$PMVL = VM - VNC$$

L'importance des PMVL réside dans le fait qu'elles peuvent être utilisées pour évaluer la performance de vos investissements. Si vos PMVL sont positives, cela signifie que l'actif a permis de réaliser des bénéfices. Si elles sont négatives, cela indique que l'actif a perdu de la valeur.

Compte tenu de la liaison évidente entre la VM, la VNC et les PMVL, nous avons décidé de ne pas prendre la VNC parmi nos variables.

2.3.2 Critère d'agrégation des actifs

Les actifs sont agrégés selon les critères suivant :

- **Code S2 de l'actif**

Il s'agit d'un code qui permet d'identifier précisément la nature de l'actif. Les actifs les plus présents dans le portefeuille étudié sont les suivants :

— ACTION_OCDE : Action appartenant à l'OCDE [1]

— ACTION_AUT : Action n'appartenant pas à l'OCDE

— OPCVM_TRANSPA : portefeuille de valeurs mobilières transparisées [2]

— OBSTRUC_OCDE : Produits structurés appartenant à l'OCDE

— TFETAT_EEA : Produits taux fixes État EEA (European Economic Area)

— TFAUT : Produits taux fixe hors État et hors hypothécaire et service publique

— IMMO : Investissement en immobilier direct, en forêt, en part de Société civile immobilière (SCI)

Le schéma ci-dessous représente une répartition du portefeuille suivant le code S2 des actifs :

1. OCDE : Organisation de Coopération et de Développement Économique constitué notamment de la France, des États_Unis, du Japon, de l'Allemagne, du Royaume-Unis,etc...

2. La transparisation des actifs est la reconstitution ligne à ligne des fonds détenus dans un portefeuille de valeurs mobilières

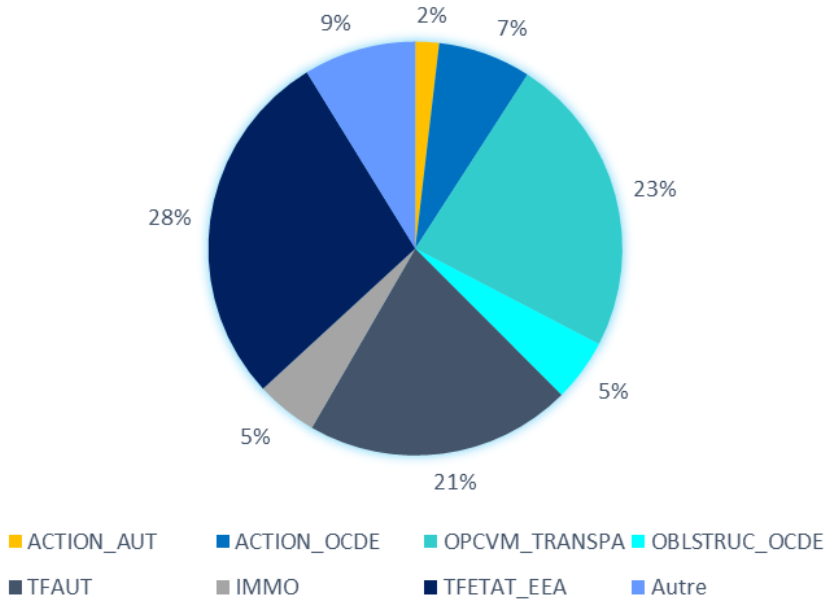

FIGURE 4.3 – Répartition de l'actif par code S2

- **Le type d'instrument**

Il s'agit de la nature d'actif. Le portefeuille est majoritairement constitué des instruments suivants :

— Les instruments de taux

— La Trésorerie

— Les actions

— Les actions protégées

— Les INC (actifs Intangibles)

Ci-dessous un graphe représentant la répartition du portefeuille par type d'instrument :

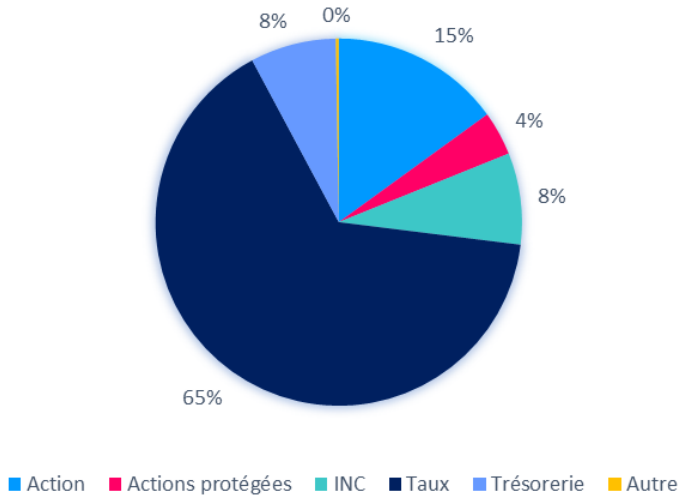

FIGURE 4.4 – Répartition de l'actif par type d'instrument

- **Notation financière de l'actif**

Il s'agit de l'évaluation d'un actif, en fonction du risque associé. La qualité d'un actif est donc importante lors de la gestion du risque puisqu'elle a un impact profond sur la rentabilité et la liquidité. Il peut s'agir des perspectives de remboursements dans le cas d'une obligation. Il s'agit donc d'un facteur clé à l'estimation du risque d'actif et à la projection de la rentabilité.

Les notations utilisées proviennent sont effectuées par les analystes au sein de la société d'assurance. Les notations vont du triple A(AAA) pour les entreprises les plus sures à C pour celles qui sont en défaut. Notre portefeuille est majoritairement constitué des actifs de type :

— SN : Actif avec zéro risque de faillite.
— AAA : Actif de première qualité, avec un risque extrêmement faible
— AA : Actif de haute qualité avec risque très faible
— A : Actif de haute qualité avec risque faible
— BBB : Actif de qualité moyenne avec risque modéré
— B : Actif avec un risque élevé en cas d'évolution défavorable des conditions commerciales ou économiques.

Après avoir agrégé l'ensemble des actifs selon ces critères, nous avons pu ressortir 8 principaux groupes. Notre base de données contient 8 variables pour chaque indicateurs, soit 24 variables au total. Dans le cas de la VBOURS par exemple, il s'agit des variables suivantes :

— VBOURS_IMMO_INC_SN
— VBOURS_OPCVM_TRANSPA_Action_SN
— VBOURS_OPCVM_TRANSPA_Taux_A
— VBOURS_TFAUT_Taux_AA
— VBOURS_TFAUT_Taux_A
— VBOURS_TFAUT_Taux_BBB
— VBOURS_TFETAT_EEA_Taux_AA
— VBOURS_TFETAT_EEA_Taux_BBB

VBOURS_IMMO_INC_SN représente alors la valeur boursière des actifs immobiliers intangibles avec un classement financier SN dans le portefeuille.
Nous avons également décider de rajouter à notre base de données la valeur boursière totale et les PMVL totales du portefeuille.

Suite à la récolte de nos données, nous disposons donc d'une base de données qui contient 87 variables.

3 preprocessing

Le preprocessing est une étape importante dans le processus de création d'un modèle de Machine Learning réussi. C'est un processus qui consiste à nettoyer et préparer les données brutes pour l'analyse. Le preprocessing est principalement composé de trois étapes : la manipulation et la transformation

des données, le nettoyage des données et la sélection des variables.

La manipulation et la transformation des données consistent à réorganiser et à modifier les données pour qu'elles soient plus adaptées à l'analyse. Cela peut inclure la fusion de plusieurs ensembles de données comme nous l'avons fait avec les variables du GSE, l'ajout ou la suppression de colonnes, la normalisation, la discrétisation, la transformation des variables. Nous aborderons par la suite la normalisation des variables.

Les données brutes peuvent contenir des erreurs, des doublons, des valeurs manquantes, des valeurs aberrantes ainsi que du bruit. Toutes ces erreurs peuvent fausser le modèle en limitant sa précision ou en le faisant apprendre de manière incorrecte. C'est pourquoi le nettoyage des données est un nécessaire.

La sélection des variables consiste à choisir les variables les plus pertinentes pour le modèle final. Cela peut être utile pour réduire la complexité du modèle et améliorer l'interprétabilité. La sélection des caractéristiques peut se faire de manière manuelle ou en utilisant des techniques de sélection automatique. Les techniques automatiques incluent l'analyse de corrélation, ou les modules feature importance des modèles utilisés.
Dans cette partie, nous présenterons une analyse des corrélations entre nos variables explicatives. Nous présenterons les modules feature importance dans le chapitre 5.

3.1 Normalisation des variables

La normalisation est un procédé courant en Machine Learning qui consiste à donner à nos variables le même ordre de grandeur, en les faisant rentrer dans l'intervalle [0,1] . Le but d'avoir un tel intervalle restreint est de réduire l'espace de variation des valeurs des variables et par conséquent réduire l'effet des outliers.

Notre base ne contient pas d'outlier mais les différentes variables ont des ordres de grandeurs très éloignés. Certaines variables sont des taux donc de l'ordre de 0,01 tandis que d'autres sont des valeurs boursières de l'ordre du million. Afin d'éviter que les variables d'un faible ordre de grandeurs ne soient considérées comme négligeables, nous devons toutes les ramener au même ordre à l'aide d'une normalisation qui ne perdra pas l'information contenue dans chaque donnée.

La normalisation se fait grâce à la formule suivante :

$$X_{normalisé} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Avec,

— X_{min} la plus petite valeur des variables observée pour une ligne donnée

— X_{max} la valeur qu'on cherche à normaliser

Cette opération nous permet de transformer toutes nos variables pour leur donner le même ordre de grandeur. La prochaine étape après cela consiste à observer les différentes liaisons qu'il existe entre nos variables.

3.2 Corrélations entre variables

3.2.1 Définitions

Comme nous l'avons déjà expliqué dans la descriptions de nos modèles, certaines dépendances entre les variables explicatives sont susceptibles d'affaiblir les performances du modèle. Il nous faut donc découvrir et identifier ces corrélations. Il existe de nombreux coefficients de corrélation pour observer les liaisons entre deux types de données. Le tableau ci-dessous regroupe les différents coefficients de corrélations en fonction leur utilisation.

Indice de mesure de la dépendance	Entre deux variables catégorielles	Entre deux variables continues	Entre une variable continue et une variable catégorielle
Corrélation de Pearson		X	
Tau de Kendall	X		
Corrélation de Spearmn		X	X
Khi 2	X		
ANNOVA			X

TABLE 4.5 – Techniques de mesure de corrélation entre deux variables

Notre base ne contenant que des variables continues, nous avons décidé de nous intéresser uniquement au coefficient de corrélation de Pearson qui s'applique dans le cas de deux variables continues.

- **Le coefficient de corrélation de Pearson**

C'est un coefficient compris entre -1 et 1 qui calcule les relations ou corrélations linéaires linéaire entre deux variables X^a et X^b. Plus il est proche de 1 en valeur absolue, plus les variables sont corrélées. Plus il est proche de 0, plus les variables sont indépendante sur le plan affine. Il se calcule comme suit :

$$\rho(X^a, Y^a) = \frac{\sigma_{X^a}, \sigma_{X^b}}{\sigma_{X^a}\sigma_{X^b}}, \ \forall a, b \in [1, p]$$

Avec,

$$\sigma_{X^a}, \sigma_{X^b} = Cov(X^a, X^b) = \mathbb{E}[(X^a - \mathbb{E}[X^a]).(X^b - \mathbb{E}[X^b])]$$

$$\sqrt{\mathbb{E}\left[(X^a - \mathbb{E}[X^b])^2\right]} = \sqrt{\mathbb{E}[(X^a)^2] - \mathbb{E}[X^a]^2}$$

On parle de Corrélation positive (le coefficient es compris dans l'intervalle]0 : 1]) lorsque les deux variables évoluent dans le même sens. La variable X_a grandit lorsque la variabe X_b augmente. On parle de corrélation négative (le coefficient es compris dans l'intervalle $[-1 : 0[$) lorsque les deux variables évoluent en sens inverse, la première variable baisse lorsque la seconde augmente et inversement.

3.3 Corrélations dans notre base de données

Au vu du grand nombre de variables que nous avons à notre disposition, nous avons décidé de ne présenter que la moyenne 1 à 5 ans pour les indicateurs économiques et une seule combinaison d'actifs pour les variables de l'actif.
L'image ci-dessous nous montre une heatmap des corrélations sur la sélection de nos variables explicatives :

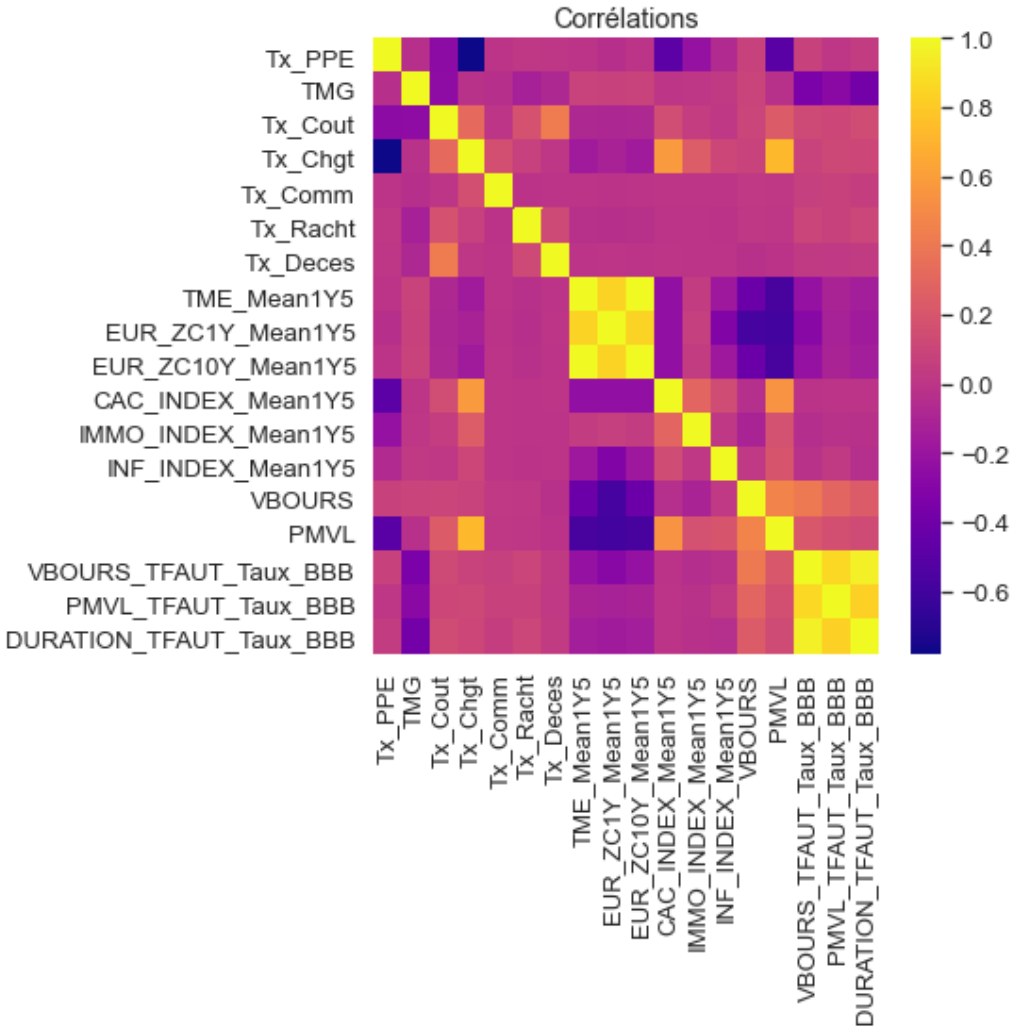

FIGURE 4.5 – Corrélation entre les variables explicatives

Plus le carré est jaune, plus la corrélation positive est forte entre deux variables. Plus le carré est violet, plus la corrélation négative est forte entre deux variables.
On observe 2 principales zones de fortes corrélations positives :

- **Corrélation entre les variables de l'actif**

On observe une corrélation forte entre les différents indicateurs du même type d'actif. Cela n'est pas surprenant vu que les PMVL d'un actif augmente si sa VBOURS augmente et inversement. On constate notamment que la duration a une forte corrélation avec les deux variables. Pour le moment, nous décidons de toutes les conserver avec réserve. Des tests supplémentaires dans les étapes à venir nous indiqueront si cela est vraiment nécessaire ou pas.

- **Corrélations entre les variables de marché**

Les secondes fortes corrélations qu'on observe sur notre figure sont celles entre le TME et les Zero Coupon. Cela n'est pas non plus étonnant. Nous avons mené des analyses supplémentaires et avons constaté le même effet de corrélation est observé pour les autres moyennes et variances. Pour chaque type d'agrégation d'indicateur économique, on observe une corrélation d'au moins 90% entre le TME et les Zero Coupon. Au vu de cette forte corrélation et des nombreuses variables économiques dont nous disposons, nous avons décidé de supprimer toutes les variables de Zero Coupon et de conserver le TME.

Le nombre de variables présentes dans notre base de données s'élève désormais à 69.

3.4 Séparation de la base de données

Pour finir, notre base de données est séparée en deux parties, pour les raisons détaillées dans le chapitre 2.
Nous avons fait le choix de la séparer en parts de ¾ et ¼ . Les 75% de la base serviront pour la phase d'apprentissage et les 25% restants pour la phase de test. Cette sélection est faite de façon aléatoire et homogène. Ainsi, la base d'apprentissage contient approximativement le même pourcentage (75%) pour chaque simulation, chaque sensibilité et chaque choc. Il en est de même pour la base de test.

Chapitre 5

Application et résultats

1 Sélection des variables explicatives et hyperparamétrisation

A présent que la construction de notre base de données est terminée, il est temps pour nous de passer aux différentes étapes de modélisations. Dans un premier temps nous présenterons les résultats et ensuite des méthodes d'amélioration parmi lesquelles la méthode par sous variables.

La première étape de notre processus consiste à lancer un premier modèle à l'aide de notre de base d'apprentissage. Cela nous permettra d'identifier les variables les plus utiles pour le modèle et de supprimer les variables non nécessaires.
Une fois les variables sélectionnées, nous pourrons passer à l'étape de la sélection des paramètres à l'aide du Gridsearch.

1.1 Sélection des variables explicatives

Afin de repérer les variables les plus utilisées au modèle, nous allons nous servir des modules « feature importance » des modèles XGBoost et random forest. En effet, comme dit plus haut, ces modèles offrent la possibilité de classer chaque variables selon son pouvoir explicatif. On pourra ainsi identifier quelle variable explicative est déterminante dans la prédiction de notre variable cible. Il convient toutefois d'interpréter les résultats du « feature importance » avec précaution car ils peuvent parfois être trompeurs ; ne correspondant pas à la compréhension et aux attendus métiers. Cette étape nous permettra néanmoins d'apprécier le choix des variables en prenant en considération un avis d'expert.

Chacun des 3 modèles que nous avons sélectionné nous donne la possibilité d'avoir un regard sur les variables qui ont le plus impacté le calcul de notre variable cible.

Le modèle XGBoost offre la possibilité de classer les variables selon 3 métriques : « Gain », « Coverage »,et « Weight ». Le modèle RandomForest offre uniquement la possibilité de classer les variables selon la métrique « Weight ».

En ce qui concerne la régression LASSO, elle ne possède pas de modules feature importance mais il y'a tout de même un moyen de savoir si des variables ont été déterminantes ou pas. En effet, le LASSO es un modèle efficace lorsqu'il s'agit de travailler avec un nombre important de variables, vu que celui-ci met à «zéro » les coefficients des variables qu'il juge inutiles au modèle. Nous allons donc observer les coefficients LASSO de nos variables pour savoir lesquelles ont été décisives lors du calcul de la VIF.

- **Métrique Weight des modèles RF et XGB**

Cette valeur représente le nombre de fois où une variable est utilisée pour diviser les données à travers l'ensemble des arbres. Elle renvoie les résultats suivants :

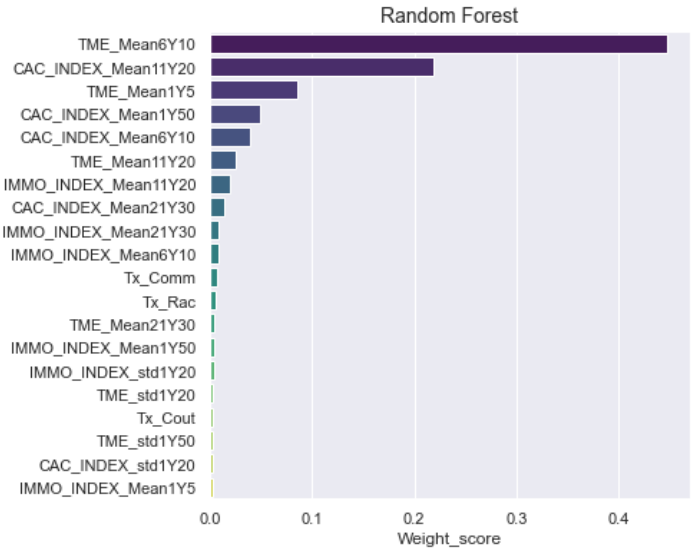

FIGURE 5.1 – Feature importance random forest

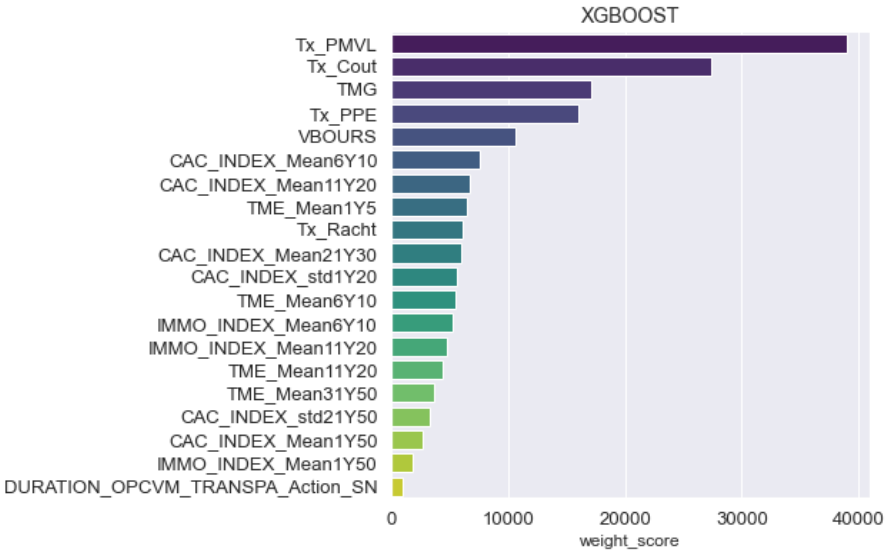

FIGURE 5.2 – Feature importance Weight XGB

On constate que cette métrique fait davantage ressortir les variables de marché pour le random forest et les variables contractuelles et du passif pour le XGBoost. Cela signifie donc que les modèles auront tendance à se servir de ces variables pour la division des données entre les nœuds.

Les variables de marché ressortent aussi beaucoup dans le modèle XGBoost. D'un côté, cela n'est pas très surprenant dans la mesure où celles-ci sont bien plus nombreuses que les autres variables. Mais si leur présence n'était dû qu'à leur nombre, alors on verrait aussi les variables de l'actif (PMVL, Duration et VBOURS par catégories d'actif) qui elles aussi sont très nombreuses. Étant donné que ce n'est pas le cas, on peut déjà penser que ces variables n'ont pas un grand impact sur le modèle. Observons ce que nous restituent les autres métriques pour en savoir plus.

- **Métriques Coverage et Gain du modèle XGB**

La métrique Gain mesure le gain sur chaque nœud, représentant la contribution de la variable sélec-

tionnée, les variables ayant les valeurs les plus importantes sont alors celles qui réduisent le plus la fonction de coût sur l'ensemble des nœuds.
La métrique Coverage représente le nombre de fois où une variables est utilisée pour diviser les données à travers l'ensemble des arbres, pondéré par le nombre de données.

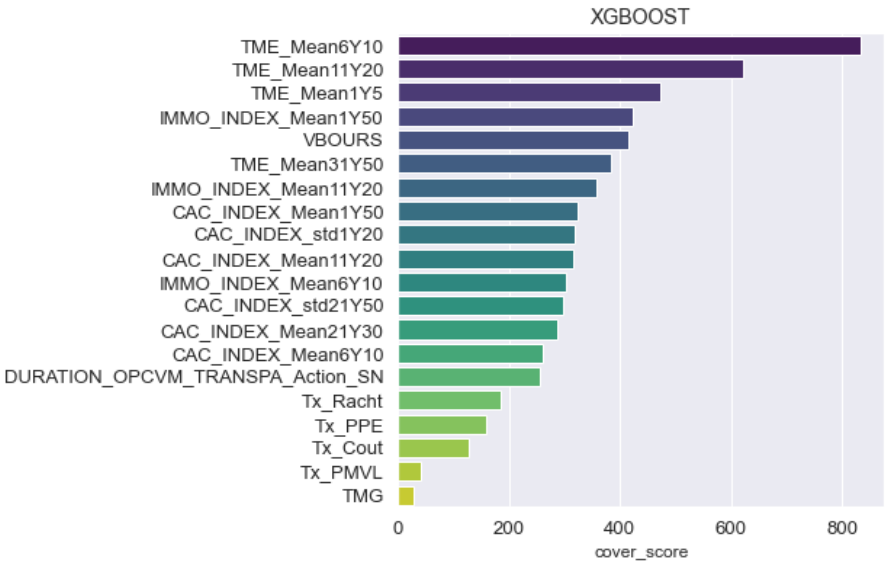

FIGURE 5.3 – Feature importance Cover XGB

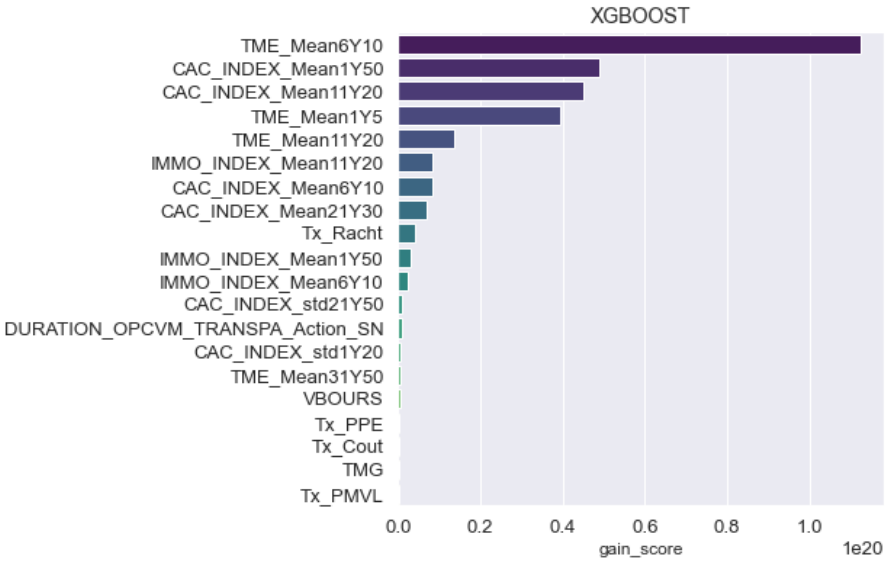

FIGURE 5.4 – Feature importance Gain XGB

En observant les résultats de ces deux métriques, on constate que les variables du passif ressortent moins et on observe principalement des variables économiques.

- **Coefficients de régression du LASSO**

Ci-dessous les 20 coefficients LASSO les plus élevés suivi des 20 coefficients nuls.

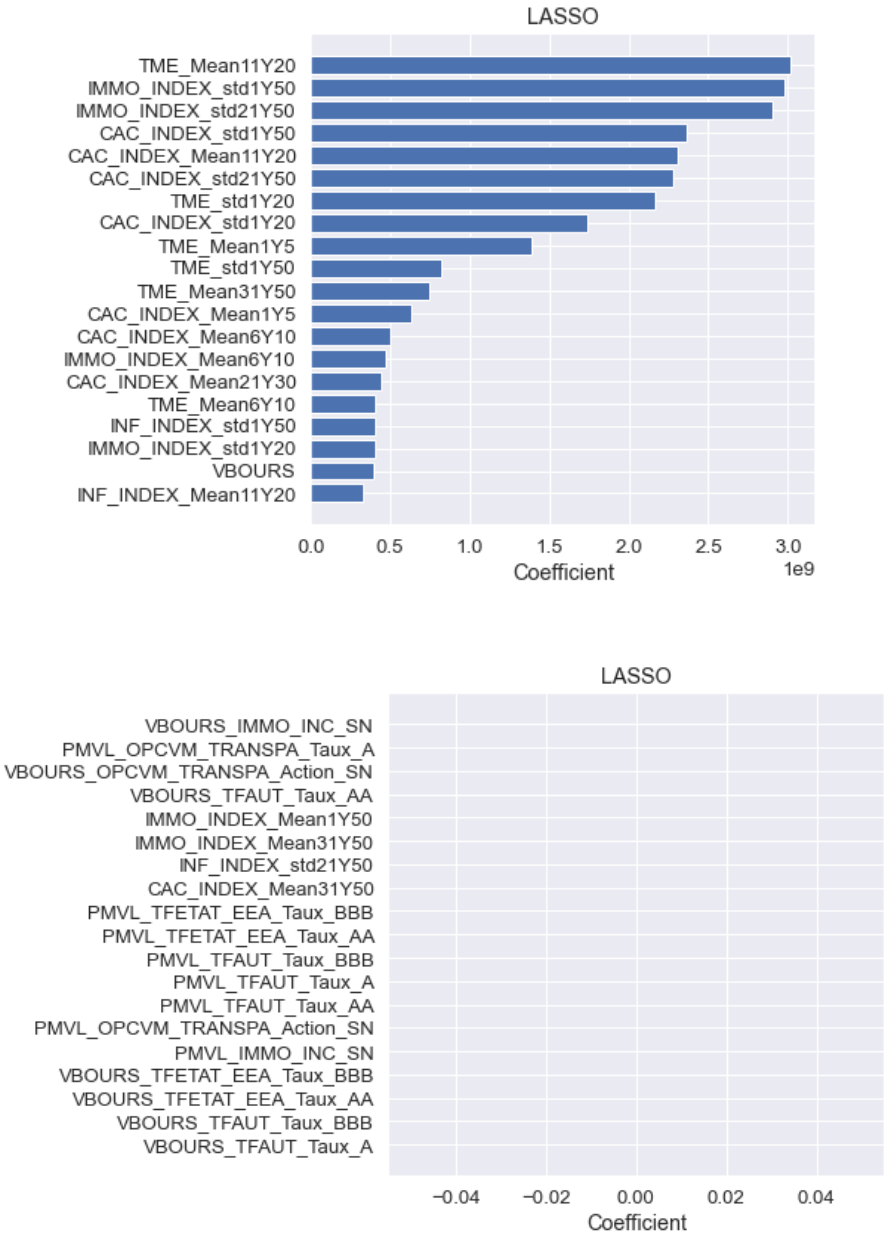

FIGURE 5.5 – Coefficients du LASSO

Les variables ont été réduites à la même échelle avant le lancement du modèle (voir dans le chapitre 4, la normalisation des variables). Nous pouvons donc être certains que les valeurs élevées des coefficients des variables de marché n'est pas dû à un problème d'échelle. D'ailleurs, le fait que la VBOURS ait un coefficient plus élevé que celui de l'inflation nous le confirme bien.

Sur la deuxième image, on observe les variables ayant un coefficient LASSO nul, ce qui signifie que ces variables n'ont pas été utilisées pour le modèle. L'ensemble des variables non utilisées est composé essentiellement des variables des PMVL et VBOURS par catégories d'actifs. Ces variables ne sont pas non plus ressorties comme étant importantes dans le classement du XGB et du RF. Et puisque nous avions déjà observé une corrélation entre la duration, la VBOURS, et les PMVL pour chaque groupe d'actif, nous avons décidé de supprimer toutes les variables PMVL et VBOURS par classe d'actifs.

Nous ne gardons que les PMVL et VBOURS totales, ainsi que les durations par groupe d'actif. Ce choix est fait par simplification, nous pouvons néanmoins revoir sa pertinence et analyser les implica-

tions.

Le nombre initial de nos variables explicatives s'élevait à 87. Suite à toutes les étapes de sur le tri des variables (corrélation et feature importance), notre base comporte désormais 53 variables explicatives.

1.2 hyperparamétrisations

La prochaine étape après la sélection des variable est la sélection des paramètres, pour nos modèles. A l'aide de notre base d'apprentissage sur laquelle on a sélectionné nos variables, nous utilisons le module GridSearch. C'est une étape très coûteuse en temps mais qui nous permettra d'obtenir de meilleurs résultats.

Les paramètres sélectionnés pour le random forest sont les suivant :

— n_estimators : 1000
— max_depth : 25
— min_sample_split : 4
— min_sample_leaf : 2
— Criterion : mse

Les paramètres sélectionnées pour le XGBoost sont les suivant :

— n_estimators : 1000
— learning_rate : 0.05
— max_depth : 10
— min_child_weight : 1
— min_sample_leaf : 2

Nous pouvons maintenant appliquer nos modèles paramétrés sur notre base d'apprentissage et observer les performances sur la base de test.

2 Premiers résultats

A présent que nous avons fait une sélection définitive de nos variables, et identifié les meilleurs paramètres pour nos modèles, nous pouvons mettre en place notre premier modèle achevé.

Nous avons retenue de calibrer nos modèles sur la base des environnements Closing, Closing_Tx_M50bps , Closing_Tx_P50bps, sans prise en compte des sensibilités croisés et ce afin d'avoir une évaluation dans un second temps de l'impact d'enrichissement de notre base d'apprentissage sur la performance de nos modèles.

Les scores de performances seront observés uniquement sur la base de test, c'est-à-dire une base que le modèle n'a pas utilisé en apprentissage. Nous n'avons pas non plus utilisé cette base pour la

calibrage des paramètres. C'est une base totalement inconnue au modèle. On pourra donc juger de la capacité des différents modèles à se généraliser sur de nouvelles données.

- **Résultats à une maille globale**

Une première présentation consiste à évaluer nos métriques de performances sur l'ensemble des données de la base de test. Nous faisons dans la suite des analyses plus approfondis via des zooms par type d'axe de simulation et par choc S2.
Le tableau ci-dessous présente les résultats obtenus sur l'ensemble des données de la base de Test.

Modèle	R2	RMSE (En M)	MRE	MAE (En M)
LASSO	0,75	1 850	180%	1423
RF	0,92	545	47%	238
XGB	0,99	289	15,7%	85

TABLE 5.1 – Résultats à une maille globale I

Les premiers résultats nous montrent que le modèle XGBoost est meilleur que les autres, du point de vue de nos 4 indicateurs. Il possède un R2 très élevé de 0,99, donc le modèle parvient à correctement deviner la distribution de notre variable. Son erreur relative absolue moyenne de 15,7% signifie que pour chaque observation prédite, le modèle commet une erreur d'environ 15,7% en plus ou en moins sur l'Equity. Cette déviation se retranscrit pas un écart de 85M € plus ou moins.

Le modèle RF arrive en deuxième position, avec un R2 moins bon de 0,89 et une RMSE de 545M. En dernière position, on a le LASSO, qui présente une RMSE de 1 850 M et une MRE de 180%. Le modèle ne semble pas adapté pour cet exercice.
Nous allons maintenant continuer d'observer les résultats sous d'autres angles.

- **Comparaison des densités**

Nous allons maintenant comparer les densités de chaque modèle avec la densité théorique afin de confirmer l'analyse de résultats. Ci-dessous on pourra voir en bleu la densité de la VIF théorique et en violet la densité de la VIF prédite par chacun de nos modèles.

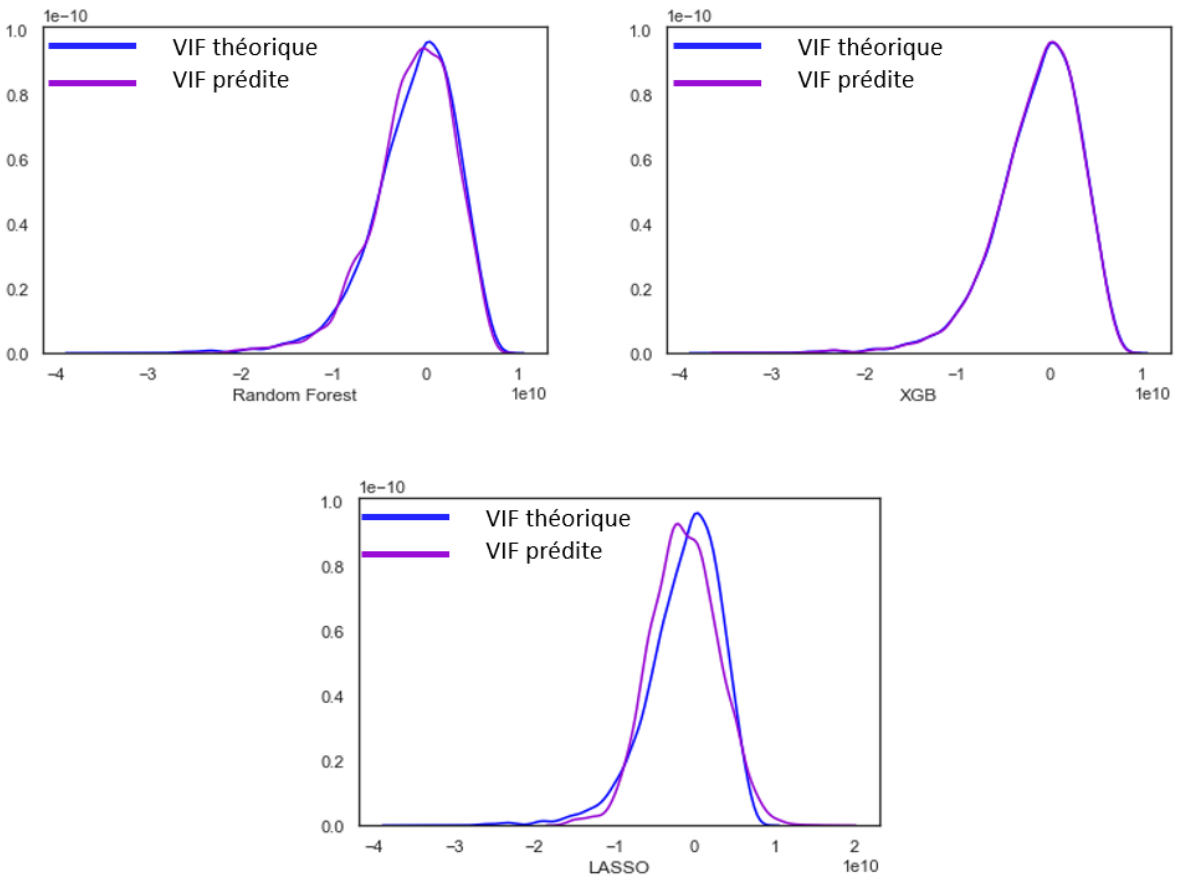

FIGURE 5.6 – Comparaison des densités I

En terme de classement des modèles, les différentes densités reflètent bien ce qu'on a pu observer à l'aide du tableau d'erreur précédent. Les analyses graphiques montrent que le XGBoost est celui qui se rapproche le mieux de la VIF théorique, suivi du random forest et du LASSO. On peut voir que la densité du modèle LASSO est assez éloignée de la variable réelle, les deux courbes n'ont pas la même structure.

- **Résultats à une maille fine**

Nous présentons dans la suite un zoom sur les résultats à une maille Closing et par type de choc SCR. Une présentation des résultats à une maille Closing_Tx_P50bps et Closing_Tx_M50bps est fournie en annexe. Nous porterons également une attention particulière à la valeur du scénario Central. Ci-dessous les résultats obtenus :

TypeChocSCR	R2 LASSO	R2 RF	R2 XGB	RE LASSO	RE RF	RE XGB
Central	0,81	0,99	0,99	11,2%	15,4%	6,1%
ChocActionType1	0,82	0,97	0,99	-10,5%	-17,6%	-6,6%
ChocActionType2	0,80	0,99	0,99	0,7%	-1,5%	0,9%
ChocBaisseRachat	0,59	0,94	0,98	28,8%	-10,0%	-0,2%
ChocBaisseTaux	0,79	0,96	0,99	-6,0%	-1,9%	-0,6%
ChocFrais	0,78	0,97	0,99	-9,0%	-19,2%	-2,4%
ChocHausseRachat	0,86	0,95	0,99	4,9%	46,5%	21,8%
ChocHausseTaux	0,76	0,82	0,92	-13,5%	-31,9%	2,3%
ChocImmobilier	0,79	0,97	0,99	-19,9%	-12,4%	-3,0%
ChocLongevite	0,71	0,96	0,99	5,7%	4,5%	-2,8%
ChocMortalite	0,85	0,99	0,99	2,3%	10,9%	0,7%

TABLE 5.2 – Résulats à une maille Closing et par choc I

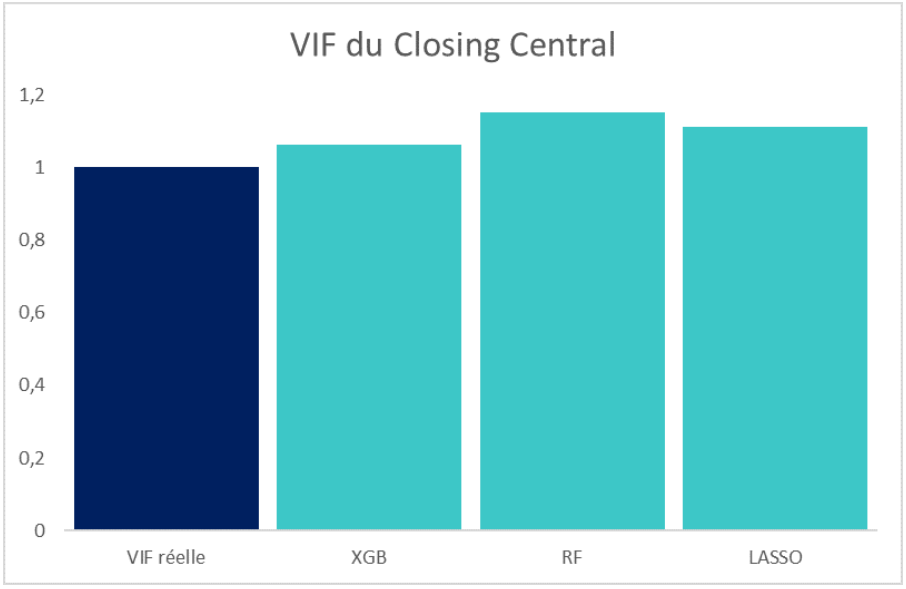

FIGURE 5.7 – Prédiction de la VIF centrale I

RE représente l'erreur relative sur le type de choc observé.

On constate que nos modèles ont tendance à surestimer la VIF centrale, le XGBoost un peu moins que les autres.

De façon générale, le XGBoost offre à nouveau des meilleurs résultats. L'ordre des modèles varient beaucoup en fonction des chocs observés mais on peut clairement voir que le modèle LASSO a tendance à avoir de meilleurs résultats que le modèle random forest cette fois, contrairement à ce qu'on pouvait observer sur nos autres métriques.

En effet, ce tableau présente les résultats de nos modèles uniquement sur le Closing. En observant les autres environnements (Closing_Tx_P50bps et Closing_Tx_M50bps), on constate que le LASSO présente des résultats nettement moins bons (écart allant souvent au-delà de 100%) tandis que le RF reste stable.
Si le LASSO semble être en mesure de prédire le Closing, il n'est pas généralisable aux environnements sur lesquels on a effectué des variations, ce qui ne nous arrange pas.

En effet le LASSO est une méthode assez simple reposant sur une hypothèse de relations linéaires entre les variables pour expliquer notre variable cible VIF. Cette méthode semble peu pertinente pour apprécier la complexité sous-jacente de la modélisation de la VIF. Les résultats XGBoost et RF mettent ainsi en exergue l'apport des approches ensemblistes permettant d'obtenir un modèle robuste pouvant capturer les effets non linéaires du phénomène modélisé.

On observe dans l'ensemble des erreurs relatives peu satisfaisantes. en effet, en parcourant l'ensemble des chocs SCR, on constate que les chocs Hausse Rachat sont moins bien prédits que les autres avec des écarts absolues qui varient entre 4,9% et 46,5%. Nous estimons que cela est dû au fait que la base ne dispose pas d'assez d'observations pour capter la variation des variable décisive pour ce choc.

Nous présentons dans la suite les résultats obtenues suite à l'enrichissement de notre base d'apprentissage.

3 Résultats suite à l'enrichissement de la base d'apprentissage

Suite aux résultats peu satisfaisants fournis par les modèles précédents, nous avons décidé de procéder à une augmentation de la base de données, coté sensibilités effectuées. Pour cela, nous avons effectué les mêmes travaux de récupération des données que précédemment. Nous avons généré de nouvelles sensibilités croisées (comme détaillé dans la section 1 du chapitre 3) et récolté les mêmes variables que précédemment. La nouvelle base de données a donc la même structure et le même nombre de variables que la précédente.

Notre nouvelle base compte 33 simulations, 11 chocs SCR, soit un total de 363000 observations.
Bien entendu, un découpage de données est également effectué afin de séparer la base d'apprentissage et la base Test.

Nous avons de nouveau fait un classement des variables les plus importantes avec les feature importance et avons obtenu les mêmes résultats que précédemment. Nous avons donc conservé la même sélection de variables pour les modèles.

En ce qui concerne le module d'hyperparamétrisation à l'aide du Gridsearch, nous n'avons pas pu l'effectuer car il était trop coûteux en temps. Nous avons donc décidé d'utiliser les même paramètres que sur notre première base. Nous allons observer les résultats en utilisant les mêmes méthodes que

précédemment. Ci-dessous les résultats obtenus :

Modèle	R2	RMSE (en M)	MRE	MAE (En M)
LASSO	0,83	1 702	180%	1 337
RF	0,99	83	9,7%	58
XGB	0,99	75	7,2%	48

TABLE 5.3 – Résultats à une maille globale II

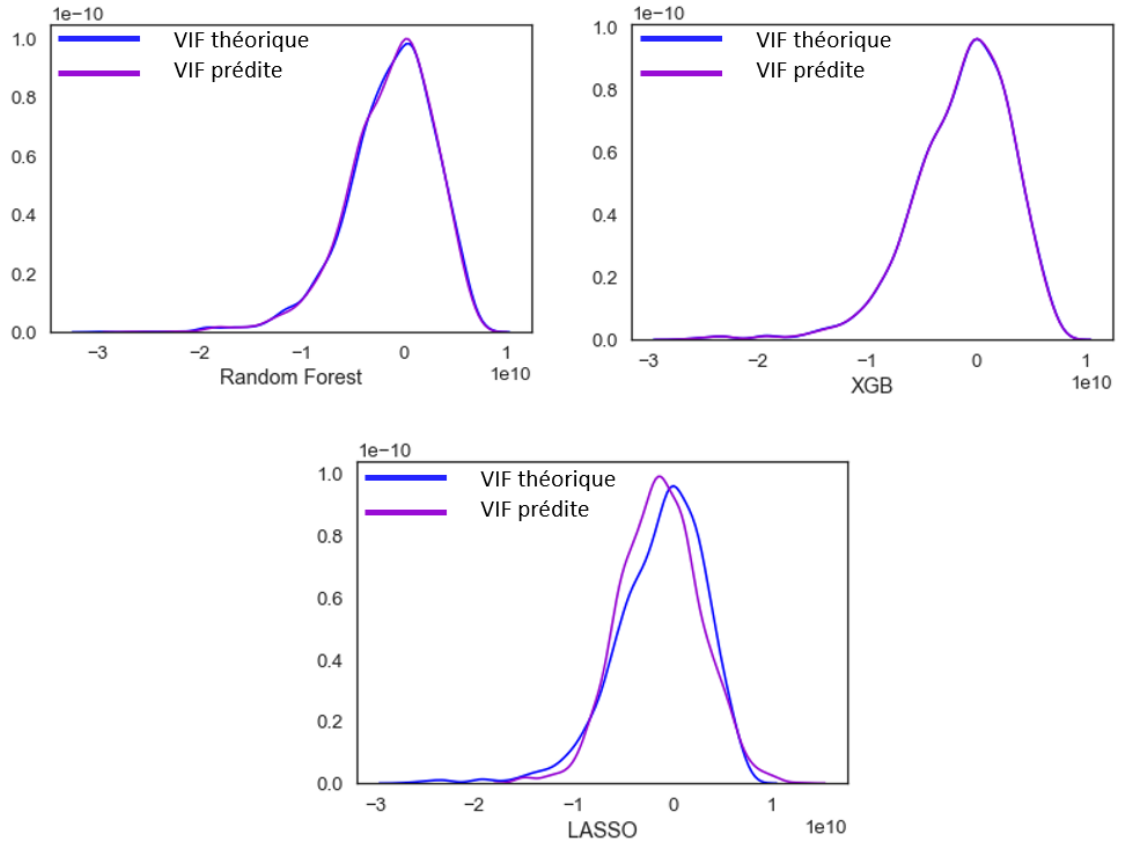

FIGURE 5.8 – Comparaison des densités II

On constate que chacun de nos modèles s'est amélioré, en particulier le random forest qui est passé d'un RMSE de 545 M à une RMSE de 83 M. Sa densité prédite s'est également beaucoup rapprochée de la densité théorique de la VIF. Toutefois, le XGBoost semble être restée le meilleur, avec une RMSE de 75M et une MRE de 7,2%. Sa densité semble aussi être plus proche de la densité théorique que celle du RF.

Observons maintenant les résultats à la maille fine. Ci-dessous les résultats obtenus :

TypeChocSCR	R2 LASSO	R2 RF	R2 XGB	RE LASSO	RE RF	RE XGB
Central	0,83	0,99	0,99	6,0%	2,5%	2,3%
ChocActionType1	0,79	0,99	0,99	-10,3%	-0,2%	0,0%
ChocActionType2	0,84	0,99	0,99	-10,4%	0,1%	0,1%
ChocBaisseRachat	0,50	0,99	0,99	10,0%	3,1%	2,9%
ChocBaisseTaux	0,79	0,99	0,99	-8,1%	2,5%	-2,7%
ChocFrais	0,81	0,99	0,99	-10,2%	-0,8%	-0,2%
ChocHausseRachat	0,87	0,99	0,99	1,7%	0,0%	-0,1%
ChocHausseTaux	0,77	0,99	0,99	-3,6%	-0,2%	-0,2%
ChocImmobilier	0,81	0,99	0,99	-13,0%	-0,6%	-0,6%
ChocLongevite	0,80	0,99	0,99	17,2%	0,0%	-0,4%
ChocMortalite	0,83	0,99	0,99	8,6%	0,3%	0,0%

TABLE 5.4 – Résulats à une maille Closing et par choc II

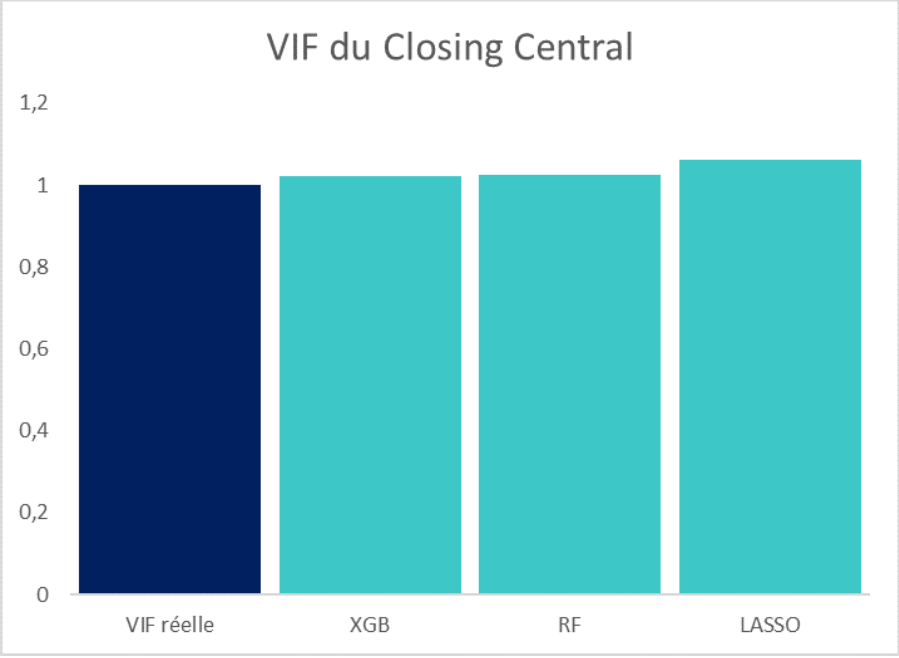

FIGURE 5.9 – Prédiction de la VIF central II

L'erreur relative (RE) du RF pour la prédiction de la VIF centrale est passée de 15,4% à 2,5%. Son erreur est désormais très proche de celle du XGBoost qui vaut 2,3%. On constate d'ailleurs que sur l'ensemble des chocs SCR, les deux modèles sont au coude à coude avec des erreurs très proches. Le modèle LASSO s'est également amélioré, passant d'un score de 10% à 6% sur le Central.

Enfin, on constate que le choc Hausse Rachat qui n'étaient pas correctement capté par les modèles

est désormais beaucoup mieux prédits. Toutes ses améliorations démontrent à quel point l'enrichissement de la base de données est une étape cruciale.
Même si les résultats se sont largement améliorés, il est important de noter que l'agrandissement de la base de données implique un fort coût en temps de calcul.

Dans la suite de notre étude, nous nous limiterons au meilleur modèle qui s'avère être le XGBoost.

4 Sélection des observations

Nous avons constaté que l'augmentation de plusieurs données avaient énormément contribué à l'amélioration de nos résultats, notamment dans le cas du random forest qui est passé d'une erreur relative de 15% à 2% pour le Closing Central. Il ne fait donc aucun doute que les observations du modèle jouent un grand rôle dans la performance, tout comme les variables.

En partant de ce fait, on propose d'approfondir les analyses sur l'impact des nouvelles observations afin d'identifier les lignes qui ont contribué à améliorer ou dégrader les prédictions et ce dans un but de mieux apprécié les résultats de nos modèles. Dans cette section nous allons faire une sélection des observations (les lignes) pour améliorer notre modèle. Compte tenu du temps que cette opération nécessite, nous avons décidé de le faire uniquement pour le modèle qui a montré les meilleurs résultats, c'est-à-dire le XGBoost.

Bien que le XGBoost possède un module qui nous permet de déterminer les variables les plus importantes, il n'existe pas à ce jour de module permettant de déterminer quelles observations ont été les plus pertinentes. De façon générales, plus il y'a d'observation et mieux c'est. Plus le nombres d'observations est grand, plus le modèle pourra apprendre. Seulement, il se pourrait que certaines observations faussent le modèle. On parle souvent de valeurs extrêmes mais dans notre cas, on ne peut pas utiliser ce terme.

En effet, l'ensemble des données récoltées proviennent de simulations cour rament effectué et sont donc correctes. Toutes fois, notre base de données n'est pas forcément apte à prédire les valeurs relatives à ces simulations en raison d'un manque potentiel de certaines variables.

Supposons maintenant par exemple que le choc S2 frais était prédit alors que nous ne disposons pas dans notre base de variables sur les frais. Le modèle essaierait de prédire la valeur de la VIF et se tromperait à coup sûr vu qu'il ne dispose pas de toutes les informations nécessaires. Il utiliserait alors une formule fausse pour se rapprocher de la valeur recherchée et cette formule serait ensuite généralisée aux autres observations, ce qui rajouterait une erreur à notre modèle. C'est pour cette raison que nous avons jugé intéressant de faire une sélection des observations.

Bien qu'il nous soit impossible de savoir quelles observations ont été déterminantes, nous sommes en mesure de savoir à quelles sensibilités et chocs SCR notre base correspond le mieux. Pour cela, deux méthodes s'offrent à nous :

- **Observation des résultats par type de sensibilité**

La première consiste à observer les résultats du Closing pour chacune de nos 32 autres simulations (qui sont des sensibilités). On pourra donc voir celles que le modèle est moins apte à prédire que les autres et se débarrasser de ces dernières.
L'observation des résultats montre que les sensibilités sur la PPE offrent de moins bons résultats que les autres, avec une erreur relative variant entre 11% et 26%

- **Mise en place d'un algorithme de sélection des observations**

La deuxième option consiste à monter par nous même un algorithme de sélection des observations. Cette algorithme va sélectionner de façon itérative les observations qui rapportent un plus au modèle. Pour le faire, il va appliquer le modèle et rajouter progressivement des observations, en observant si ces observations améliorent le score du modèle ou pas.

La mise en place de cet algorithme nécessite le lancement du modèle d'un nombre de fois égal aux observations (363 000 fois ici). Cette opération étant bien trop coûteuse en temps, nous avons décidé de remplacer les observations par des simulations, et de faire des sélections par sensibilités. Cela revient à lancer le modèle 32 fois seulement.
Notre algorithme se décline de la façon suivante :

Algorithme : Sélection des sensibilités pour le modèle

Données : Ensembles des sensibilités disponibles S
XGB = modèle XGB
score = score du modèle actuel

Initialisation L = {Closing}, liste qui gardera les meilleures sensibilités

Pour sensi_i compris dans la liste des sensibilités S **faire**
    L' = L + sensi_i
    Effectuer un modèle xgb sur la base de L'
    score_i = MAE de ce modèle
    Si score_i ≤ score **faire** :
        score = score_i
        Rajouter la sensi_i qui réduit l'erreur à L
Fin **Fin**

Sortie : Retourner L la liste des meilleures sensibilités

Cet algorithme part donc du Closing comme sensibilités principale et ensuite itère chaque nouvelle sensibilité et sauvegarde dans la liste L contenant uniquement les sensibilités qui rapportent un plus au modèle. Les résultats de l'algorithme nous ont montré ce qu'on redoutait déjà plus haut : les sensibilités sur la PPE ont été exclues car elles font baisser le résultat du modèle.

Ces simulations feront l'objet d'investigations pour mieux rationaliser ces effets. En effet la base d'apprentissage pourrait être enrichie avec des variables complémentaires reflétant des règles de gestion de la PPE.

Pour la suite de nos travaux, nous décidons donc de nous débarrasser de toutes ces sensibilités (6 simulations au total sur la PPE donc 66000 observations).
Nous relançons à nouveau notre modèle de XGBoost, qui avait été le plus performant des 3 jusqu'ici, en excluant les sensibilités sur la PPE de nos variables. Nous obtenons les résultats suivant :

Modèle	R2	RMSE (en M)	MRE	MAE (en M)
XGB	0,99	35	3,2%	20

TABLE 5.5 – Résultats à une maille globale III

Type choc SCR	R2 XGB	Ecart XGB (en M)	RE XGB
Central	0,99	23	1,6%

TABLE 5.6 – Résulats à une maille Closing et par choc III

On constate que nos résultats ce sont encore améliorés. L'écart passe de 2,3% à 1,6%. Le fait que notre base données soit moins grande a permis d'aller le temps de calcul. Cet étape nous donc permis de gagner en performance et en temps de calcul.

5 Approche par décomposition en sous variables

A ce stade de nos travaux, nous savons désormais quelles sensibilités, quelles variables et quel modèle sont les plus pertinents pour la prédiction de notre variable cible. Désormais, nous pouvons tester la seconde méthode de projection que nous avons appelée «approche par décomposition en sous variables».
Pour rappel, cette approche consiste à prédire individuellement les 3 principales constituantes de la VIF puis à les agréger. En effet, la VIF peut être décomposée comme suit :

$$VIF = Prélèvements - Commission - Couts$$

Nous allons donc prédire individuellement chacune de ses sous variables, en utilisant les techniques et méthodes qui nous ont mené au meilleur résultat comme discuté précédemment. Étant donné que le XGBoost nous a donné de meilleures performances dans la prédiction directe, nous l'avons réutilisé, pour construire 3 différents modèles pour chacune de nos sous variables.

Nous n'avons pas fait pour chaque sous variable une étape de sélection de variable explicative pertinente. Nous avons utilisé pour toutes les sous variables les mêmes variables qui ont servi à la calibration du modèle de la VIF. Cela a du sens car la VIF est en réalité une combinaison linéaire de nos sous variables. D'un point de vue théorique, si nos différentes variables explicatives sont reliée à la VIF alors elles sont également reliées à nos sous variables qui sont en réalité une décomposition de la VIF.

D'un point de vue expert, cela a également du sens car pour chacune de ses sous variables, il existe une ou plusieurs variables explicatives qui lui sont étroitement liées et permettraient de l'expliquer. Par exemple, la variable explicative « Taux de Coût » est fortement liée à la sous variable « Coûts ». On observe les résultats suivant :

Modèle	R2	RMSE (en M)	MRE	MAE (en M)
VIF XGB	0,99	34,8	2,6%	20,3
VIF Ss variables	0,99	28,0	1,4%	15,4
Coûts	0,99	5,7	0,2%	3,3
Commissions	0,99	8,5	0,6%	4,6
Prélèvements	0,99	13,4	0,7%	7,9

TABLE 5.7 – Résultats à une maille globale IV

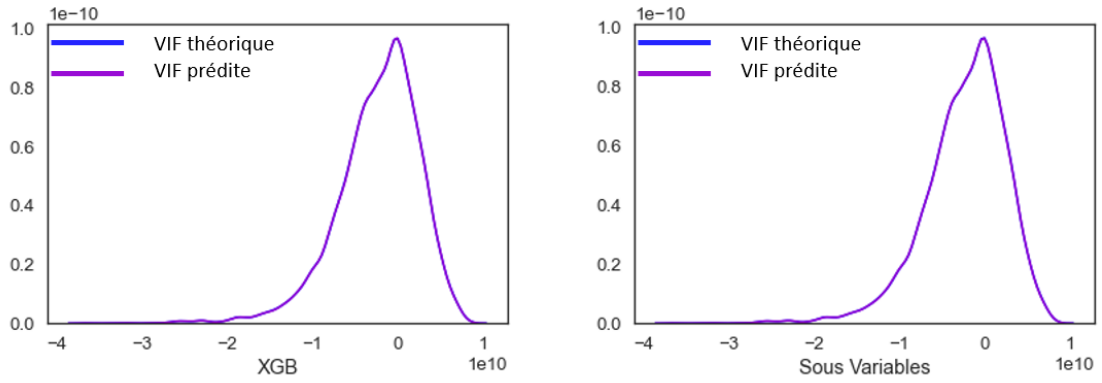

FIGURE 5.10 – Comparaison des densités III

TypeChocSCR	R2 XGB	R2 ssVar	RE XGB	RE SS VAR
Central	0,99	0,99	1,6%	1,0%

TABLE 5.8 – Résulats à une maille Closing et par choc IV

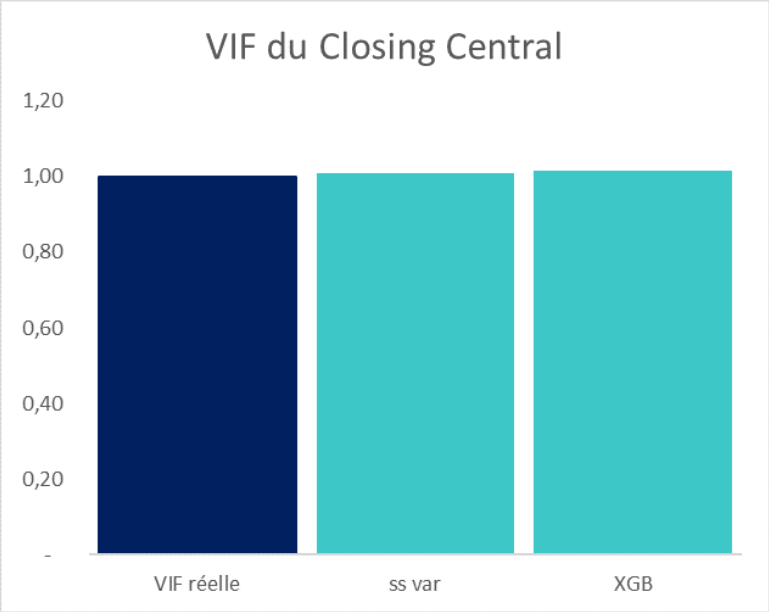

FIGURE 5.11 – Prédiction de la VIF centrale III

Le fait de séparer notre variable cible en 3 variables moins complexes nous a permis d'améliorer légèrement notre score, passant de 1,6% à 1,0%. On peut observer que le modèle arrive à capter parfaitement la densité de la variable théorique et le R2 (0,99) montre que le modèle permet d'appréhender la complexité cachée dans le calcul de la VIF.
La méthode par sous variable avec le modèle XGBoost est donc celle qui offre de meilleurs résultats.

6 Calculs de SCR

Nous allons présenter dans cette partie l'utilisation de notre modèle pour le calcul de SCR. Le SCR d'un choc peut être calculé comme la différence entre la VIF pour un scénario choqué et la VIF du scénario central.
Le calcul des SCR pour les chocs (sous modules de risques) disponibles de notre base de données nous donne les résultats suivant :

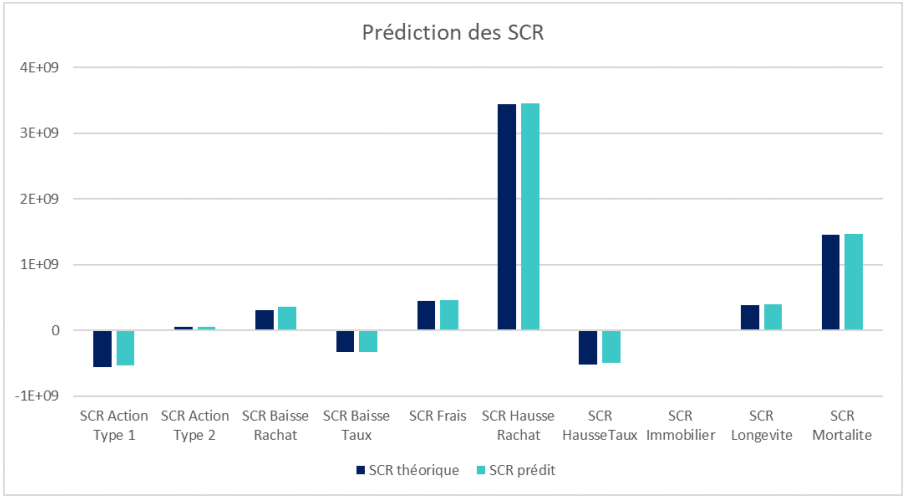

FIGURE 5.12 – Prédiction du SCR

On peut observer que les prédictions du modèle sont à chaque fois très rapprochées du SCR observé. Le modèle arrive à capter les variations des différents SCR présentés.
Compte tenu des sous-modules étant à notre disposition, nous allons maintenant calculer les SCR des modules marché et souscription vie. Comme nous l'avons expliqué dans le chapitre 1, le SCR d'un module résulte de l'agrégation des SCR des différents sous modules correspondant à l'aide d'une matrice de corrélation du sous module correspondant.

$$SCR_m = \sqrt{\sum_{i,j} \rho_{(i,j)} C_i C_j}$$

Avec,

— SCR_m le SCR du module correspondant

— C_i et C_j les SCR des sous modules i et j

— $\rho_{(i,j)}$ la matrice de corrélation des sous modules considérés

Nous allons utiliser les matrices de corrélations suivante :

Matrice de corrélation	Actions	Taux	Immobilier
Actions	1	0,5	0,75
Taux	0,5	1	0,5
Immobilier	0,75	0,5	1

Table 5.9 – Matrice de corrélation pour le SCR marché

Matrice de corrélation	Mortalité	Longévité	Rachat	Frais
Mortalité	1	-0,25	0	0,25
Longévité	-0,25	1	0,25	0,25
Rachat	0	0,25	1	0,5
Frais	0,25	0,25	0,5	1

Table 5.10 – Matrice de corrélation pour le SCR de souscription vie

Après nos calcul, on obtient les résultats suivants :

SCR calculé	Valeur théorique (en M)	Valeur prédite (en M)	Ecart (en M)	Erreur relative
SCR marché	60,7	60,0	0,69	1,1%
SCR vie	4 076	4 113	-37,23	-0,9%

TABLE 5.11 – Résultats du calcul des SCR par module

On observe une erreur de 1,1% sur le SCR marché et -0,9% sur le risque de souscription vie. Les résultats obtenus nous paraissent satisfaisants.
Le modèle que nous avons mis en place nous paraît relativement fiable en matière de résultats, bien que le modèle reste difficile à interpréter.

En effet l'observation des résultats ne compense pas l'effet « boite noire » du modèle. Afin de pouvoir nous fier à notre modèle, il nous faut plus de détails, plus d'explications sur la façon dont il calcule ses résultats et les variables qu'il utilise.
Nous allons donc expliquer notre modèle à l'aide d'un module complémentaire d'explicabilité type SHAP.

7 Transparence du modèle

7.1 Définition

La communication et l'interprétation des résultats d'un modèle Machine Learning reste souvent délicate et difficile, en raison de la complexité des règles générées par le modèle et afférant aux résultats. C'est pourquoi, on parle souvent d'effet « boite noire » des modèles en Machine Learning. L'ensemble des décisions prises par le modèle et des calculs effectuées pour estimer la variable cible ne sont pas facilement compréhensibles et disponibles.
Les algorithmes de transparence entrent alors en jeu dans le but de proposer une solution à ce problème.

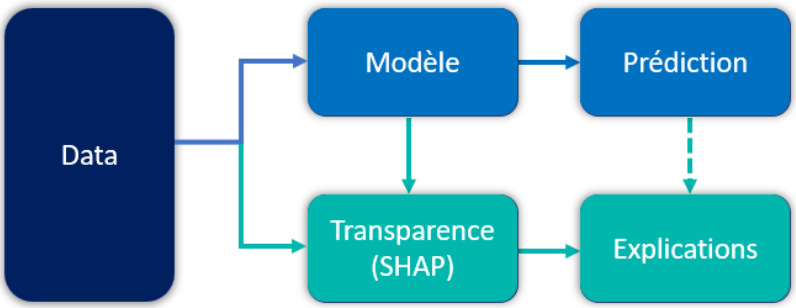

FIGURE 5.13 – Algorithme de transparence du modèle

Les algorithmes de transparence sont des algorithmes de ML spécifiques qui visent à rendre les décisions prises par les modèles plus compréhensibles et explicables. Ils ajoutent une surcouche au modèle entraîné afin d'analyser les prédictions et fournir des éléments facilitant la compréhension du

modèle, permettant ainsi d'améliorer la confiance envers le modèle.
Les algorithmes de transparence peuvent être classés suivant 4 principaux types :

Les algorithmes d'interprétabilité
Les algorithmes d'interprétabilité sont conçus pour rendre les décisions prises par un Modèle de Machine Learning plus compréhensibles pour les humains. Ils fournissent des visualisations et des explications en langage naturel pour aider les utilisateurs à comprendre comment le modèle prend des décisions.

Les algorithmes de perturbation
Les algorithmes de perturbation consistent à modifier ou supprimer des parties de l'ensemble de données d'entraînement, de manière à évaluer comment le modèle de ML pourrait réagir à des données qu'il n'a pas vues auparavant. Cela peut aider à identifier les biais et les vulnérabilités des modèles.

Les algorithmes de validation
Les algorithmes de validation sont utilisés pour évaluer les performances des modèles de ML et les comparer aux données d'entraînement. Ils peuvent aider à identifier les modèles qui sous-performent ou qui ont des biais importants.

Les algorithmes de transparence de la décision
Les algorithmes de transparence de la décision permettent aux utilisateurs de voir comment les décisions sont prises par le modèle de ML. Ils peuvent fournir des informations sur les données et les variables qui ont été pris en compte, ainsi que sur la façon dont elles ont été pondérés.

Notre objectif étant de fournir une explication simple qui reste fidèle dans une certaine mesure aux décisions réellement prises par le modèle, nous avons décidé de nous orienter vers un algorithme d'interprétabilité, appelé SHAP. Ce choix est motivé par le fait que c'est l'une des méthodes les plus populaires et pertinentes dans la littérature.

7.2 SHAP

L'algorithme de transparence SHAP, (SHapley Additive exPlanations), est un algorithme qui vise à décomposer de manière rationnelle la prédiction d'un modèle complexe de Machine Learning. Cette méthode de décomposition est basée sur le concept de valeur de Shapley, qui est souvent utilisée dans la théorie des jeux coopératifs.

Quand un modèle réalise une prédiction, intuitivement nous savons que chaque variable ne joue pas le même rôle et même que certaines n'ont quasiment aucun impact tandis que d'autres en ont un grand sur la décision prise par le modèle. Nous l'avons dalleurs constaté en observant l'importance des variables fournis par les modèles plus haut.

L'objectif de la valeur de Shapley est donc de quantifier le rôle de chaque variable dans la décision finale du modèle. Pour comprendre comment cette valeur fonctionne, nous allons commencer par présenter le principe de la valeur de Shapley dans la théorie des jeux.

7.2.1 Valeur de Shapley

L'indice de Shapley est un concept important en théorie des jeux coopératifs. Il a été développé par Lloyd Shapley en 1953 et est utilisé pour déterminer la contribution équitable de chaque joueur dans un jeu coopératif où les gains sont partagés entre tous les joueurs.

Dans un jeu coopératif, chaque joueur collabore avec un autre pour atteindre un objectif commun. À la fin du jeu, les gains sont partagés entre tous les joueurs. Cependant, il est souvent difficile de déterminer la contribution équitable de chaque joueur à la victoire du groupe. L'indice de Shapley résout ce problème en attribuant une valeur à chaque joueur en fonction de la contribution qu'il a apportée à la victoire du groupe. Cette valeur est calculée en considérant toutes les combinaisons possibles de joueurs qui peuvent contribuer à la victoire du groupe.

Considérons une jeu de coopération J , on note :

— Le 2-uplet $J = (\{1, ..., p\}, v)$ notre jeu de coopération

— p le nombre de joueurs et $P = \{1, ..., p\}$ l'ensemble des joueurs

— $v : S(P) \to \mathbb{R}$ une fonction caractéristique telle que $v(\varnothing) = 0$. Il s'agit du gain total.

— $S(p)$ l'ensemble des sous ensembles de P

— $S \in S(P)$ un sous ensemble de joueurs

— $v : S(P) \to \mathbb{R}$ une fonction caractéristique telle que $v(\varnothing) = 0$

— ϕ_i l'importance ou la participation du joueur i dans le gain total

Lloyd a réussi à montrer que la valeur de Shapley respecte les 4 propriétés suivantes :

1. **Efficacité**
   $\sum_{i=1}^{p} \phi_i(v) = v(\{1, ..., p\})$

2. **Symétrie**
   Soit $(i, j) \in \{1, ..., p\}^2$ un couple de joueurs. Si $\forall S \in S(P \backslash \{i, j\})$, $v(S \cup i) = v(S \cup j)$, alors $\phi_i(v) = \phi_j(v)$

3. **Facilité**
   Soit $i \in P$ un joueur. Si $\forall S \in S(P \backslash \{i\})$, $v(S \cup \{i\}) = 0$, alors $\phi_i(v) = 0$

4. **Additivité**
   Pour tout jeu v ω, on a :
   — $\phi(v + \omega) = \phi(v) + \phi(\omega)$
   — Avec $\forall S \in S(P)$, $(v + \omega)(S) = v(S) + \omega(S)$

La valeur de shapley qui distribue le gain total $v(P)$ pour chaque joueur i est :

$$\phi_i(v) = \sum_{S \in S(P \backslash \{i\})} \frac{|S|!(p - |S| - 1)!}{p!} (v(S \cup \{i\}) - v(S))$$

7.2.2 La valeur de Shapley appliquée à l'interprétabilité des modèles

L'idée derrière SHAP est d'utiliser la valeur de Shapley pour expliquer les prédictions faites par un modèle complexe de Machine Learning.
La fonction de gain sera alors la fonction de prédiction et la valeur de Shapley va représenter une mesure de l'importance relative de chaque variable dans une prédiction.

Pour adapter la formule de la théorie des jeux à contexte, on va noter : :

— p le nombre de variable du modèle et P l'ensemble des variables

— f la fonction de prédiction qui remplace la fonction des gains

— i la i-ème variable

— $S(p)$ l'ensemble des sous ensembles de P

— $S \in S(P)$ un sous ensemble de variables

La valeur de Shapley se calcule comme suit :

$$\phi_i = \sum_{S \in S(P \setminus \{i\})} \frac{|S|!(p - |S| - 1)!}{p!} (f_t(S \cup \{i\}) - f_t(S))$$

L'approche SHAP est additive, donc une prédiction peut être écrite comme la somme des différents effets des variables (valeur de Shapley υ_i) ajoutée la valeur de base υ_0. La valeur de base étant la moyenne de toutes les prédictions du dataset :

$$\hat{f}(X) = Y_p red = \upsilon_0 + \sum_{i=1}^{p} \upsilon_i z_i$$

Avec,

— $\hat{Y}$ la valeur prédite par le modèle

— υ_O la moyenne de toutes les prédictions

— $z \in 0, 1$ prend la valeur 1 lorsque la variable est utilisée et 0 sinon

Grâce à la valeur de Shap, on peut donc déterminer l'effet des différentes variables d'une prédiction pour un modèle qui explique l'écart de cette prédiction par rapport à la valeur de base.

7.2.3 Application du SHAP à notre modèle

L'algorithme SHAP peut être utilisé grâce au package SHAP disponible sous Python. L'enjeu opérationnel de l'estimation des indices de Shapley est qu'ils nécessitent de simuler les n ! permutations de l'ensemble des n variables d'entrée, ce qui représente un énorme coût en temps.
Le SHAP permet d'interpréter le modèle suivant deux angles axes d'analyse : un axe global et un axe local.

- **SHAP Observation globale**

L'axe global permet d'expliquer de façon général les principales règles de décision du modèle. Comme nous l'avons expliqué, la valeur de Shapley est calculée en mesurant les gains ou les pertes de performance du modèle lorsqu'une variable est ajoutée ou retirée de l'ensemble des variables. De cette façon,

l'algorithme peut identifier les variables qui ont le plus d'impact sur la prédiction finale.

La vision sous l'axe global permet donc d'observer au global les variables ayant le plus d'impact sur la prédiction et de quantifier ces impacts.
Ci-dessous les résultats obtenus :

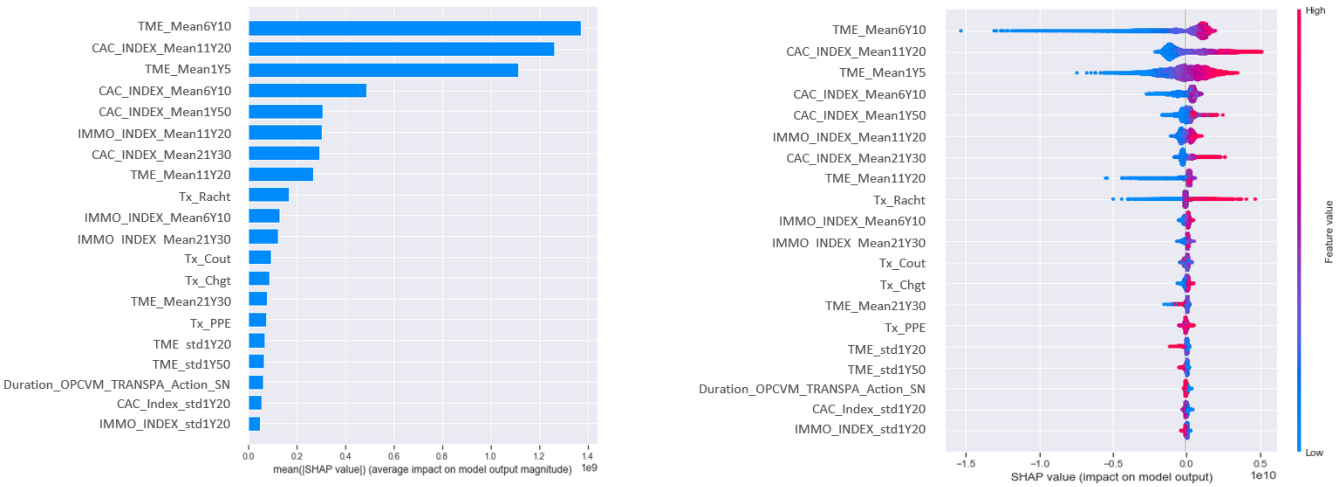

FIGURE 5.14 – SHAP Observation globale

Le graphique de gauche montre les variables les plus importantes classées par ordre décroissant et celui de droite leurs contributions à l'explication de la VIF. SHAP indique également comment chaque variable affecte la variable dépendante VIF.
Les variables les plus importantes qui découlent du modèle XGBoost sont principalement le niveau des TME et du CAC_Index. On note qu'un niveau élevé du CAC Index a un impact fort et positive sur la prédiction de la VIF. Ceci nous semble conforme à l'attendu dans la mesure où la richesse du portefeuille est cohérente par rapport à la performance de la poche actions.

- **Observation locale**

L'axe local permet d'expliquer le calcul de la variable cible pour une observation donné.
Nous avons mentionné plus haut que SHAP permettait d'écrire chaque prédiction comme la somme entre la moyenne et des valeurs de Shapley.

L'approche locale permet donc d'observer comment est-ce que la prise en compte des variables contribue à faire diverger la valeur d'une prédiction de la moyenne des prédictions.
Ci-dessous les résultats obtenus :

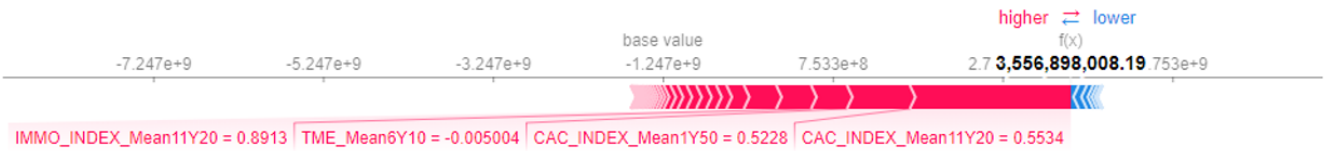

FIGURE 5.15 – SHAP Observation locale

Ce présente les variables qui contribuent à impacter la prédiction à partir de la valeur de base. On note que les variables qui poussent la prédiction à la hausse (vers la droite) sont affichées en rouge, et celles qui poussent la prédiction à la baisse sont en bleu. Dans le cas de notre observation, les principales variables impactant le résultat final sont des variables économiques.
Dans le cas de notre exemple, on constate que la valeur prédite s'éloigne à la hausse de la valeur moyenne en raison des variables CAC, TME et IMMO.

8 Conclusion

Dans le but d'obtenir un modèle fiable et généralisable, nous avons dû passer par plusieurs étapes depuis la sélection des variables jusqu'à l'application de la méthode « sous variables ». Nous avons noté de nombreuses façons d'améliorer notre modèle.

Le graphique ci-dessous montre les différentes évolutions qu'on a observé sur le modèle XGBoost, depuis le premier lancement de modèle visant à sélectionner les variables et jusqu'à la valeur finale de la méthode par sous variables :

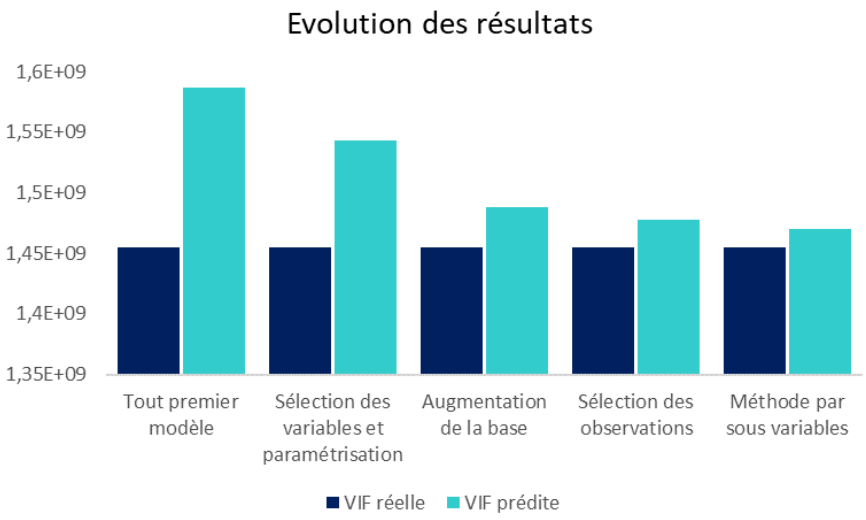

FIGURE 5.16 – Evolution des résultats au cours des travaux

Conclusion

Dans ce mémoire, nous avons essayé de prédire la valeur d'un portefeuille épargne (noté VIF) à l'aide de méthodes de Machine Learning. Notre objectif était de fournir un outil rapide et explicable qui pourrait répliquer la complexité des modèle actuariels Actif/Passif afin de mesurer la rentabilité d'un assureur.

Pour cela, nous avons été amenés à comprendre à la fois les principes cachés derrière les modèles actuariels et également ceux du Machine Learning. Cela nous a permis d'un côté de construire une base de données solide qui contiendrait alors les informations utiles et nécessaires pour la prédiction de notre variable cible. D'un autre côté, nous avons aussi pu déterminer les modèles de Machine Learning pertinents ainsi que les différentes méthodes qui pourraient être utilisées pour améliorer nos résultats.

Nous avons sélectionné 3 modèles à savoir la régression LASSO, le random forest et le XGBoost. La régression LASSO nous a apporté une approche simple et moins complexe tandis que les autres modèles étaient plus complexes mais également plus précis car capables de capter des effets non linéaires. Nous avons également utilisé 2 méthodes à savoir une approche de prédiction directe de la VIF et une approche par décomposition en sous variables de la VIF.

A la fin de notre travail, nous avons observé que le modèle XGBoost avec une approche par décomposition en sous variables était le plus performant, avec des résultats assez satisfaisants. En effet, ce dernier s'avère plus acceptable en terme de biais et de généralisation que les autres. Les analyses de transparences réalisées restent toutefois relatives à la notion de simplicité qui dépend de l'interlocuteur. Les méthodes proposées ne peuvent en aucun cas remplacer les approches actuelles mais permettent toutefois d'estimer rapidement et parfois avec précision la métrique en considération, facilitant ainsi les sensibilités nécessaires au pilotage et à la prise de décision.

Notre approche exploratoire nous a convaincu qu'il reste plusieurs pistes d'investigations à entrevoir pour améliorer les résultats. En effet, ce genre de modélisation n'est pas sans risque d'erreurs, avec des biais qui peuvent être de sources différentes on note par exemple le risque d'avis d'expert, de sélection, de données et de variables omises.
La base d'apprentissage pourrait être enrichie à l'aide de nouvelles sensibilités et on pourrait éventuellement trouver de meilleurs méthodes pour incorporer la stratégie d'investissement et les règles de gestion et de revalorisation spécifiques au portefeuille étudié.
Il pourrait être envisageable de tester d'autres modèles de Machine Learning tels que les réseaux de neurones afin d'observer leur apport.

Bibliographie

A. Buzzi. Approximation du bilan économique sous Solvabilité II via des méthodes d'apprentissage automatique et application à l'ORSA. 2017. Mémoire, Dauphine.

ACPR. Solvabilité II : principaux enseignements de la cinquième étude quantitative d'impact (QIS5). 2017

A. Guillot. Apprentissage statistique en tarification non-vie : quel avantage opérationnel ? 2015. Mémoire, ENSAE.

A.Messoussi. Application d'algorithmes de machine learning pour l'estimation du ratio de couverture d'un assureur-vie détenteur d'un produit épargne. Mémoire, ISFA. 2017.

A.S. Dalalyan. Apprentissage et data mining. ENSAE ParisTech 3ème année.

BELLINA, R. Méthodes d'apprentissage appliquées à la tarification non-vie. Mémoire, ISFA. 2014

C. Boyer. Machine Learning. Cours Master 2, ISUP. 2020.

D. Delcaillau. Contrôle et Transparence des modèles complexes en actuariat. Mémoire, ENSAE. 2019

DELOITTE. Documentation modèle et Formation interne CNP Assurances.

DREYFUS G., MARTINEZ J.M., SAMUELIDES M., GORDON M.B., BADRAN F., THIRIA S. Apprentissage statistique, Eyrolles. 2008.

E. Biernat and M. Lutz. Data Science : Fondamentaux et études de cas. Eyrolles, 2015.

E. Strumbelj. Shapley sampling values : "Explaining prediction models and individual predictions with feature contributions." Knowledge and information systems.2014.

E. Zurfluh. Utilisation du Machine Learning dans l'estimation du ratio de solvabilité d'un assureur vie et application aux Reverse Stress Tests. Mémoire, ENSAE. 2019

H. KACEM. Estimation de la valeur d'un portefeuille épargne avec une approche Machine Learning. 2020. Mémoire DSA Data Science pour l'Actuariat, IRM Institut du Risk Management. 2020.

J. Friedman. Greedy Function Approximation : A Gradient Boosting Machine. 1999.

J.Sac, M. Petit et M. Donio. FORMULE STANDARD USP, Publishroom, Paris, ISBN 979-10-236-0086-5. Guide d'aide à la réalisation des calculs Solvabilité II, Sia partners. 2020.

J. Starmer. Statquest with Josh Starmer. https ://www.youtube.com/statquest.

L. Breiman, J. Friedman, and Al. Classication and Regression Trees. 1984.

L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen. Classification and Regression Trees. The Wadsworth and Brooks-Cole statistics-probability series. Taylor Francis, 1984.

L. Breiman. Random forests. Mach. Learn., 45(1) :5–32, October 2001.

L. Shapley. A value for n-person games. Contributions to the Theory of Games. 1953.

M. Fromont. Apprentissage statistique. Cours, Ensai. 2015.

M. VELUT. Enjeux et modélisation de l'ajustement pour risque sous la forme IFRS 17. 2018. Mémoire, Université de Strasbourg, Duas.

R. Gauville. Projection du ratio de solvabilité : des méthodes de machine learning pour contourner les contraintes opérationnelles de la méthode des SdS. 2017. Mémoire, Dauphine.

S. Fortmann-Roe. Understanding the Bias-Variance Tradeoff. 2012.

T. Chen and C. Guestrin. XGBoost : A scalable Tree Boosting System. 2016.

X. Dupre. Enseignements. http ://www.xavierdupre.fr/app/ensae_teaching_cs/helpsphinx3/index.html . Cours, Ensae.

Annexe

Résultats complémentaires

Ci dessous les résultats de modélisation sur les environnements Closing_Tx_M50bps et Closing_Tx_P50bps avant l'enrichissement de la base d'apprentissage.

TypeChocSCR	R2 Lasso	R2 RF	R2 XGB	RE Lasso	RE RF	RE XGB
Central	0,77	0,82	0,99	13,8%	18,4%	2,8%
ChocActionType1	0,75	0,82	0,99	-5,8%	-12,4%	-2,5%
ChocActionType2	0,72	0,84	0,99	-7,0%	-4,6%	-0,3%
ChocBaisseRachat	0,16	0,80	0,99	-9,3%	-12,3%	-1,6%
ChocBaisseTaux	0,71	0,87	0,99	-3,7%	9,5%	1,3%
ChocFrais	0,68	0,84	0,99	-6,9%	-10,5%	-1,7%
ChocHausseRachat	0,86	0,86	0,99	8,9%	27,4%	3,9%
ChocHausseTaux	0,75	0,79	0,97	-111,0%	-11,9%	-26,1%
ChocImmobilier	0,77	0,87	0,99	-10,5%	-9,6%	-1,2%
ChocLongevite	0,72	0,87	0,99	-0,6%	7,9%	-0,1%
ChocMortalite	0,82	0,87	0,99	4,9%	9,1%	0,6%

TABLE 5.12 – Résulats à une maille Closing_Tx_M50bps et par choc

TypeChocSCR	R2 Lasso	R2 RF	R2 XGB	RE Lasso	RE RF	RE XGB
Central	0,87	0,93	0,99	36,5%	-27,8%	-15,1%
ChocActionType1	0,85	0,90	0,99	-18,3%	-31,1%	-14,0%
ChocActionType2	0,84	0,90	0,99	-12,6%	-13,9%	-8,0%
ChocBaisseRachat	0,72	0,84	0,99	-170,7%	-94,4%	3,1%
ChocBaisseTaux	0,87	0,93	0,99	82,9%	41,6%	2,5%
ChocFrais	0,84	0,87	0,99	-185,7%	61,4%	61,5%
ChocHausseRachat	0,81	0,92	0,99	66,9%	-78,7%	-7,2%
ChocHausseTaux	0,67	0,84	0,94	10,8%	-21,4%	-5,4%
ChocImmobilier	0,85	0,92	0,99	-43,3%	-22,2%	-2,5%
ChocLongevite	0,81	0,90	0,99	-522,4%	-72,8%	-30,8%
ChocMortalite	0,80	0,90	0,99	-3,9%	-7,2%	-2,5%

TABLE 5.13 – Résulats à une maille Closing_Tx_P50bps et par choc

Ci-dessous les résultats à maille fine de la modélisation des sous variables de la VIF :

TypeChocSCR	R2 Prl	R2 Comm	R2 Coûts	RE Prl	RE Comm	RE Coûts
Central	0,99	0,99	0,99	-0,8%	-0,2%	-0,1%
ChocActionType1	0,99	0,99	0,99	1,1%	0,2%	0,0%
ChocActionType2	0,99	0,99	0,99	0,2%	-0,1%	0,0%
ChocBaisseRachat	0,99	0,99	0,99	0,0%	0,3%	0,1%
ChocBaisseTaux	0,99	0,99	0,99	0,6%	0,0%	0,1%
ChocFrais	0,99	0,99	0,99	-0,2%	0,0%	-0,1%
ChocHausseRachat	0,99	0,99	0,99	0,2%	0,1%	0,2%
ChocHausseTaux	0,99	0,99	0,99	-0,2%	-0,5%	0,0%
ChocImmobilier	0,99	0,99	0,99	0,6%	0,1%	0,0%
ChocLongevite	0,99	0,99	0,99	0,0%	0,1%	0,0%
ChocMortalite	0,99	0,99	0,99	0,1%	0,0%	0,1%

TABLE 5.14 – Résulats des sous variables à une maille Closing et par choc

Table des figures

Liste des tableaux