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

Par : Ghada BEN YOUSSEF

Titre du mémoire : Elaboration d'un proxy Machine Learning du BEL à partir de sa sensibilité aux aléas du marché

Confidentialité : ☒ NON ☐OUI (Durée : ☐ 1 an ☐2 ans)

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

Membres présents du jury de la filière : Signature :	Entreprise : BNP Paribas Cardif Nom : MEHIDI Signature :
	Directeur de mémoire en entreprise Nom : MEHIDI Signature :
Membres présents du jury de l'Institut des Actuaires : Signature :	Invité : Nom : Signature :
	Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels (après expiration de l'éventuel délai de confidentialité) Signature du responsable entreprise : Signature du candidat :

Remerciements

Je remercie tout d'abord mon premier tuteur Ademola TOUKOUROU et ma deuxième tutrice Sylia MEHIDI qui ont su m'accompagner dans mon sujet parfois en cours de route et s'armer de patience afin de répondre autant que possible à tous mes questionnements.

Je remercie de même ma collègue Johanna BUAYI avec qui j'ai partagé d'agréables moments et dont les conseils et le retour d'expérience ont été déterminants dans l'aboutissement de ce mémoire. Elle a été d'un véritable soutien surtout durant les moments difficiles. Merci à mes deux collègues Montassar TAMMAR et Ines BURR.

J'adresse mes remerciements à Philippe BAUDIER le manager de notre équipe de m'avoir apporté une aide très précieuse plus d'une fois.

Je remercie également mon tuteur académique Olivier LOPEZ pour ses remarques et critiques qui m'ont guidée.

Je souhaite enfin remercier mon conjoint Ali ELLOUZE d'être la personne merveilleuse qu'il est, de me soutenir, de m'encourager durant toutes les épreuves et de me pousser sans cesse à atteindre mes ambitions. J'espère pourvoir un jour en faire de même pour toi ! Merci enfin à ma mère, Lilia Atig. Mes réussites ne sont que les fruits de ses sacrifices.

Résumé

Depuis l'entrée en vigueur le 1er janvier 2016 de la directive Solvabilité II, les assureurs ont, une fois de plus, du faire preuve d'adaptabilité face à la nouvelle réforme. Depuis, de nouvelles problématiques émergent dont le besoin d'efficience quant aux calculs des indicateurs de reporting de la réglementation.

Le BEL ou Best Estimate Liabilities est un indicateur estimant les engagements probables de l'assureur. Son calcul se fait grâce au modèle ALM de BNP Paribas Cardif. Cependant, ce processus est coûteux en temps et l'évaluation plus fréquente du BEL peut s'avérer utile pour l'assureur tant pour la prise de décision que pour se projeter et évaluer sa solvabilité (capacité à faire face à ses engagements).

L'objectif de ce mémoire est de produire un proxy du BEL pour les contrats d'épargne fonds Euro à partir des aléas du marché. Concrètement, l'étude évalue la sensibilité de l'indicateur face à la variation de diverses variables représentant la fluctuation du marché : la PMVL, la PPE et la courbe des taux.

Des chocs simples, doubles et triples sont appliqués à ces variables à partir des données 2018, 2019 et 2020. Les résultats de ces chocs sont ensuite intégrés comme inputs au modèle ALM afin de déterminer le BEL correspondant. Une base de données est, par la suite, constituée à partir des diverses valeurs d'inputs ainsi que d'autres variables non choquées et du BEL d'output.

La base composée des données 2018 et 2019 est utilisée pour des fins d'apprentissage supervisé i.e. pour calibrer divers modèles de Machine Learning aux données. Finalement, le modèle présentant la meilleure performance est choisi et validé à travers, entre autre, un Backtesting sur les données 2020.

Mots-clés : *BEL*, *proxy*, *Solvabilité II*, *Machine Learning*, *épargne*, *fonds Euro*, *PMVL*, *PPE*, *courbe des taux.*

Abstract

Since the entry into force on the 1st of January 2016 of the Solvency II directive, insurers have, once again, had to demonstrate their adaptability to the new reform. Ever since, new issues are emerging among which the need for efficiency in the calculation of regulatory reporting indicators.

The BEL or Best Estimate Liabilities is an indicator estimating the probable liabilities of the insurer. It is calculated using the BNP Paribas Cardif ALM model. However, this process is time-consuming and the more frequent assessment of the BEL could come in handy for the insurer both for decision-making and for planning and assessing his solvency (ability to meet his commitments).

The objective of this dissertation is to produce a proxy of the BEL for Euro fund savings contracts through market movements. Concretely, the study assesses the sensitivity of the indicator to the variation of various variables representing the fluctuation of the market: the PMVL, the PPE and the yield curve.

Single, double and triple shocks are applied to these variables from 2018, 2019 and 2020 data. The results of these shocks are then integrated as input to the ALM model in order to determine the corresponding BEL. A database is then created from the various input values as well as other non-shocked variables and the output BEL.

The database composed of 2018 and 2019 data is used for supervised learning purposes i.e. to calibrate various Machine Learning models to the data. Finally, the model with the best performance is chosen and validated through, among other things, Backtesting on 2020 data.

Keywords : *BEL*, *proxy*, *Solvability II*, *Machine Learning*, *savings*, *Euro fund*, *PMVL*, *PPE*, *yield curve*.

Synthèse

Problématique et contexte

Dans le cadre de la réglementation Solvabilité II, l'assureur est tenu de présenter divers indicateurs quantitatifs visant à évaluer sa capacité à faire face à ses engagements. L'un de ces indicateurs est le BEL. C'est l'estimateur probable des engagements de l'assureur selon divers scénarios économiques.

Le calcul de cet indicateur est effectué selon un processus assez complexe et coûteux en temps. Par ailleurs, il est parfois nécessaire de déterminer le BEL sur une maille de temps plus fine que celle du reporting. Cet indicateur peut s'avérer précieux dans le cadre de décisions d'investissement, offre de la visibilité à l'assureur en fonction des éventuels aléas du marché et reste à tout moment un élément représentatif de sa solvabilité.

Ce mémoire tente de résoudre, en partie, cette problématique en générant un modèle permettant d'approximer le calcul du BEL pour les contrats d'épargne du fonds Euro. Plus précisément, il s'agit d'évaluer la sensibilité de l'indicateur par rapport à diverses variables captant la volatilité du marché.

La solution apportée

Dans un contexte où le Machine Learning occupe une place centrale en analyse de données, cette méthode est choisie pour apporter une réponse à la problématique. Cette technique répond au besoin principalement quantitatif plutôt qu'interprétable du résultat.

De plus, elle présente de nombreux avantages par rapport à l'utilisation d'une formule fermée à savoir : déceler des liens éventuellement dissimulés entre les variables (complexes ou indirects)

et permettre une plus grande complexité/précision du modèle.

Concrètement, il a fallu déterminer les trois variables faisant partie du modèle existant qui reflètent au mieux le comportement du marché. Les variables principales conservées sont la PMVL, la PPE et la courbe des taux.

Naturellement, une approche classique en actuariat est adoptée : celle des stress tests. Des chocs sont appliqués aux variables explicatives et une base de données est générée à partir de ces données et des valeurs de BEL y correpondant. D'autres variables considérées comme secondaires ont aussi été intégrées sans être choquées à cette base de données pour offrir de la précision. Cette base de données sert ensuite comme input aux algorithmes d'apprentissage.

La méthodologie

Cette étude s'articule comme suit :

— Description du modèle ALM de Cardif dont un des indicateurs de sortie est le BEL. Cette description situe les variables d'intérêt ainsi que le BEL dans le modèle ;

— Création de la base de données composée des variables primaires, secondaires et du BEL à partir des données des années 2018, 2019 et 2020 ;

— Analyse préalable théorique des relations entre les variables principales-mêmes et entre ces dernières et le BEL ;

— Visualisation de ces relations dans le cas réel à partir des données et confrontation de ces résultats aux attentes de l'analyse préalable ;

— Calibration de divers modèles d'apprentissage supervisé sur la base de données ;

— Comparaison entre les performances des divers modèles et choix du modèle à conserver ;

— Résultats et validation du modèle choisi.

Création de la base de données

La base de données a été créée à partir de plusieurs types de chocs : les simples, les doubles et les triples.

Les premiers consistent à faire varier une seule des variables principales et de conserver les deux autres à leurs valeurs réelles (appelées valeurs centrales). Les chocs doubles (respectivement triples) font varier deux des (respectivement toutes les) variables principales en même temps et gardent la valeur de la troisème intacte.

Ces variables sont des inputs du modèle ALM et les modifier revient à altérer les tables d'inputs de ce dernier. Le modèle ALM de BNP Paribas Cardif est implémenté sur Prophet. Il s'agit d'un outil usuel en actuariat permettant de projeter le bilan de l'assureur sur un horizon de 40 ans dans le cas de Cardif.

Pratiquement, il prend en compte divers inputs tels que les portefeuilles d'actifs et de contrats des assurés et des scénarios économiques reflétant les états du marché (à travers la table ESG). Il génère en output plusieurs indicateurs comme le BEL ou le SCR. Différentes valeurs du BEL sont calculées en sortie, chacune correspondant à une simulation de l'évolution possible du marché. L'output gardé pour la base de données est la moyenne de toutes les réalisations du BEL.

Les chocs appliqués à la PMVL sont : $\pm 1\%,\ \pm 2\%,\ \pm 3\%,\ \pm 4\%,\ \pm 5\%,\ \pm 10\%\ et\ \pm 20\%$. Ceux de la PPE sont de $\pm 5\%,\ \pm 10\%,\ \pm 15\%,\ \pm 20\%,\ \pm 25\%,\ \pm 30\%,\ \pm 50\%,\ \pm 80\%\ et\ \pm 100\%$ alors que la courbe des taux n'a subie que deux variations : $\pm 10\%$. Les données sont celles des années 2018, 2019 et 2020.

Intégrer la courbe des taux sans transformation à la base de données n'est pas raisonnable puisqu'elle est composée de nombreux points. Une solution choisie est de la résumer en quatre valeurs à l'aide du modèle de Nelson Siegel.

Ce modèle capte la majorité de l'information contenue dans la courbe grâce à quatre paramètres : β_0 traduisant le niveau de la courbe, β_1 paramètre de pente de la courbe, β_2 reflétant la déformation ou courbure de la courbe et λ pour la vitesse de décroissance de la courbe. Ce sont les paramètres Nelson Siegel correspondant aux neuf courbes (trois scénarios pour chaque année) qui ont alimenté la base de données.

Comportement du BEL en fonction des différentes variables

Choquer la PMVL revient essentiellement à altérer la valeur de marché des actions du portefeuille d'actifs de l'assureur. Ceci affecte les produits financiers de l'assureur et donc le mécanisme de participation aux bénéfices.

La PPE, quant à elle, est une provision dotée dans le cas de bénéfices importants. Elle fait partie du mécanisme de participation aux bénéfices. L'assureur puise dans celle-ci en cas de gains insuffisants afin de servir le taux cible aux assurés pour rester compétitif et éviter les rachats dynamiques.

Enfin, la courbe des taux sert principalement à évaluer les obligations du portefeuille d'actifs de l'assureur. Choquer cette courbe déprécie ou valorise ces obligations. Comme pour la PMVL, les revenus de l'assureur sont affectés ainsi que le mécanisme de la participation aux bénéfices.

L'analyse préalable dévoile une corrélation positive entre le BEL et la PMVL et la PPE ainsi qu'une autre entre ces deux dernières. Au contraire, en raison de la formule de valorisation des obligations, la corrélation est négative entre le mouvement de la courbe des taux et celui des trois variables citées (le BEL, la PPE et la PMVL).

Une autre attente énoncée au cours de cette analyse préalable est un effet plus important sur le BEL en cas de bénéfices qu'en cas de pertes (ou de gains faibles). Cette particularité est expliquée par l'utilisation des diverses provisions (dont la PPE et certaines variables secondaires) dont dispose l'assureur dans ce dernier cas.

La visualisation du BEL en fonction des différentes valeurs des trois variables principales a confirmé parfaitement cette analyse théorique. Elle a aussi permis de souligner une relation quasi-linéaire entre le BEL et la PMVL et la PPE ainsi qu'une plus grande sensiblité du BEL par rapport aux variations de la PPE que la PMVL. L'interprétation de ceci est que la PPE est représentative du mécanisme de participation aux bénéfices dans sa globalité. Ce mécanisme intègre les bénéfices et des actions (liés à la PMVL) et des obligations (liés à la courbe des taux). Une variation de la PPE se répercuterait donc sur ces deux types de bénéfices d'où un effet plus considérable.

Modélisation du BEL à l'aide d'algorithmes Machine Learning

Une étude de la matrice de corrélation entre les variables de la base de données a relevé une importante corrélation entre la RC, la PDD et le Cash. Une seule de ces trois variables a donc été conservée : la PDD.

Le boxplot des valeurs du BEL a montré que les données sont assez homogènes sans valeurs aberrantes. La distribution du BEL souligne une légère asymétrie en faveur des grandes valeurs. Ceci confirme les attentes de l'analyse préalable privilégiant un BEL plus élevé même en cas

de faibles bénéfices.

La base de données a été scindée en deux : un bloc composé des données de 2018 et 2019 et un autre contenant les données de l'année 2020. L'apprentissage a été effectué sur le premier bloc. Le deuxième bloc a servi de base pour le Backtesting.

Le bloc composé des données de 2018 et 2019 qui a permis de calibrer les modèles a subi un brassage afin d'éviter les tendances temporelles puis divisé en deux blocs : 70% ont servi de base train (d'apprentissage) et les 30% restant de base test.

Plusieurs modèles d'apprentissage supervisé ont été implémentés. Tout d'abord, la régression linéaire a présenté une capacité prédictive assez correcte. Ensuite, deux modèles linéaires généralisés (GLM) ont été calibré sur les données : un Gaussian GLM et un Gamma GLM. Le second s'est avéré être plus performant compte tenu de l'asymétrie naturelle de sa distribution et qui a été détectée aussi au niveau de celle du BEL.

Divers modèles de type Decision Trees ont ensuite été appliqués dont les Simple Decision Trees, Adaboost, Gradient Boosting et Random Forest. Ces modèles ont tous présenté une performance exceptionnelle sur les données.

Enfin, ce mémoire détaille l'utilisation d'un algorithme de Deep Learning : les Réseaux de Neurones. Cet algorithme a présenté des prédictions globalement correctes mais assez chaotiques et médiocres par rapport à tous les autres algorithmes y compris les modèles linéaires.

Résultats et validation du modèle choisi

La comparaison de performance entre tous ces modèles a reposé sur leur performance sur la base test et sur diverses méthodes d'évaluation. L'utilisation d'indicateurs de performance fait partie de ces méthodes : le BIC, la RMSE, le score et la cross-validation. Une représentation des boxplots des erreurs relatives a aussi permis de comparer les modèles.

Le modèle choisi au final est celui du Gradient Boosting. Il a présenté la meilleure performance. Tous les modèles (dans une moindre mesure celui des Réseaux de Neurones) ont cependant été performants. Si le critère prépondérant au niveau du choix est l'interprétabilité alors le modèle GLM Gamma est optimal.

Le modèle Gradient Boosting choisi a présenté une excellente performance sur la base de

données de Backtesting, i.e. celle de l'année 2020. Ceci, en plus de l'homogénéité des données, d'une bonne performance de modèles simples (tels que la régression linéaire) et de bons résultats de cross-validation ont permis d'éliminer le risque d'overfitting. Ceci a permis de valider le modèle.

Cette étude a été effectuée sur un grand nombre de données. Ses résultats sont assez cohérents avec l'analyse théorique. Les prédictions sont excellentes et le modèle choisi a été validé plus d'une fois : sur la base de données test puis grâce au Backtesting. Les résultats sont donc fiables.

Cependant, plusieurs choix ont été effectués au niveau de cette étude dont les variables à utiliser et l'approche Machine Learning. Tous ces choix peuvent être remis en question et les altérer pourrait résulter en un meilleur proxy du BEL. Enfin, il serait intéressant d'étendre cette étude pour produire des proxies d'autres indicateurs comme le SCR, par exemple.

Synthesis

Problem and context

Under the Solvency II regulations, the insurer is required to present various quantitative indicators with the aim of assessing its ability to meet its commitments. One of these indicators is the BEL. It is the probable estimator of the insurer's commitments according to various economic scenarios.

The calculation of this indicator is carried out in a rather complex and time-consuming process. Furthermore, it is sometimes necessary to determine the BEL over a finer time scale than that of the reporting. This indicator can prove to be valuable in the context of investment decisions, offers visibility to the insurer depending on potential market variations and remains at all times a representative element of its solvency.

This thesis tries to solve, in part, this problem by generating a model allowing to approximate the calculation of the BEL for the savings contracts of the Euro fund. More precisely, its objective is evaluating the sensitivity of the indicator in relation to various variables capturing the volatility of the market.

The provided solution

In a context where Machine Learning occupies a central place in data analysis, this method is chosen to provide an answer to the problem. This technique responds to the mainly quantitative rather than interpretable need for the result.

In addition, it has many advantages compared to the use of a closed formula, namely: detecting any hidden links between the variables (complex or indirect) and allowing greater

complexity/accuracy of the model.

Concretely, it was necessary to determine the three variables forming part of the existing model that best reflect the behavior of the market. The main variables retained are the PMVL, the PPE and the yield curve.

Naturally, a classic actuarial approach is adopted: that of stress tests. Shocks are applied to the explanatory variables and a database is generated from these data and the corresponding BEL values. Other variables considered secondary were also integrated without being shocked into this database to provide precision. This database is then used as input to the learning algorithms.

Methodology

This study is structured as follows:

— Description of Cardif's ALM model as one of its output indicators is the BEL. This description locates the variables of interest as well as the BEL in the model;

— Creation of the database composed of primary and secondary variables as well as the BEL from data corresponding to the years 2018, 2019 and 2020;

— Prior theoretical analysis of the relationships between the main variables themselves and between the latter and the BEL;

— Visualization of these relationships in the real case from the data and comparison of these results with the expectations of the preliminary analysis;

— Calibration of various supervised learning models on the database;

— Comparison between the performances of the various models and choice of the model to keep;

— Results and validation of the chosen model.

Creation of the database

The database was created from several types of shocks: singles, doubles and triples.

The first consist in varying only one of the main variables and keeping the other two at their real values (called central values). Double (respectively triple) shocks vary two of (respectively all) the main variables at the same time and keep the value of the third intact.

These variables are inputs of the ALM model and modifying them amounts to altering the input tables of the latter. BNP Paribas Cardif's ALM model is implemented on Prophet. This is a common actuarial tool used to project the insurer's balance sheet over a 40-year horizon in the case of Cardif.

Practically, it takes into account various inputs such as portfolios of assets, portfolios of clients' policies and economic scenarios reflecting the states of the market (through the ESG table). It generates in output several indicators such as the BEL or the SCR. Different values of the BEL are calculated as output, each corresponding to a simulation of the possible evolution of the market. The output kept for the database is the average of all the realizations of the BEL.

The shocks applied to the PMVL are: $\pm 1\%$, $\pm 2\%$, $\pm 3\%$, $\pm 4\%$, $\pm 5\%$, $\pm 10\%$ *and* $\pm\, 20\%$. Those to the PPE are $\pm 5\%$, $\pm 10\%$, $\pm 15\%$, $\pm 20\%$, $\pm 25\%$, $\pm 30\%$, $\pm 50\%$, $\pm 80\%$ *and* $\pm\, 100\%$ while the yield curve has undergone only two variations: $\pm 10\%$. The data is for the years 2018, 2019 and 2020.

Integrating the yield curve without transformation into the database is not reasonable since it contains many points. A chosen solution is to summarize it in four values using the Nelson Siegel model.

This model captures the majority of the information contained in the curve thanks to four parameters: β_0 representing the level of the curve, β_1 parameter of the slope of the curve, β_2 reflecting the deformation or curvature of the curve and λ for the rate of decay of the curve. The Nelson Siegel parameters corresponding to the nine curves (three scenarios for each year) are the ones that integrated the database.

Behavior of the BEL according to the different variables

Shocking the PMVL essentially results in altering the market value of equites in the insurer's portfolio of assets. This affects the financial products of the insurer and therefore the profit-sharing mechanism.

The PPE, on the other hand, is a provision endowed in the event of large profits. It is part of the profit-sharing mechanism. The insurer draws on this in the event of insufficient gains in order to serve the target rate to the insured so as to remain competitive and avoid dynamic surrenders.

Finally, the yield curve is mainly used to value the insurer's asset portfolio bonds. Shocking this curve depreciates or values these bonds. Like for the PMVL, the income of the insurer is affected as well as the mechanism of profit sharing.

The preliminary analysis reveals a positive correlation between the BEL and the PMVL and the PPE as well as a similar one between the latter two. On the contrary, because of the bond valuation formula, the correlation is negative between the movement of the yield curve and that of the three variables aforementioned (the BEL, the PPE and the PMVL).

Another expectation stated during this preliminary analysis is a greater effect on the BEL in the event of profits than in the event of losses (or small gains). This particularity is explained by the use of the various provisions (including the PPE and certain secondary variables) available to the insurer in the second case.

The visualization of the BEL according to the different values of the three main variables confirms this theoretical analysis perfectly. It also highlights a quasi-linear relationship between the BEL, the PMVL and the PPE as well as a greater sensitivity of the BEL to variations in the PPE than the PMVL. The interpretation of this is that the PPE is representative of the profit-sharing mechanism as a whole. This mechanism integrates earnings of equities (linked to the PMVL) and bonds (linked to the yield curve). A change in the PPE would therefore have repercussions on these two types of benefits, resulting in a more considerable effect.

BEL modeling with Machine Learning Algorithms

A study of the correlation matrix between the variables of the database reveals a significant correlation between the RC, the PDD and the Cash variable. Only one of these three variables has therefore been kept: the PDD.

The boxplot of the BEL values shows that the data is fairly homogeneous with no outliers. The distribution of the BEL highlights a slight asymmetry in favor of large values. This confirms the expectations of the prior analysis favoring a higher BEL even in the event of weak profits.

The database was split into two: a block composed of data from 2018 and 2019 and another containing data from the year 2020. Training was performed on the first block. The second block served as the basis for Backtesting.

The block composed of data from 2018 and 2019 which allowed the models to be calibrated was shuffled in order to avoid temporal trends and then divided into two other blocks: 70% served as the train database (for learning) and the remaining 30% as the test database.

Several supervised learning models have been implemented. First of all, linear regression presented a rather correct predictive capacity. Then, two generalized linear models (GLM) were calibrated on the data: a Gaussian GLM and a Gamma GLM. The second turned out to be more efficient given the natural asymmetry of its distribution, which was also detected for the BEL.

Various Decision Tree-type models were then applied, including Simple Decision Trees, Adaboost, Gradient Boosting and Random Forest. These models all showed outstanding performance on the data.

Finally, this thesis details the use of a Deep Learning algorithm: Neural Networks. This algorithm presented globally correct but quite chaotic and poor predictions compared to all other algorithms including linear models.

Results and validation of the chosen model

The comparison between all these models was based on their performance on the test dataset and on various evaluation methods. The use of performance indicators is one of these methods: the BIC, the RMSE, the score and the cross-validation. A representation of the boxplots of the relative errors also made it possible to compare the models.

The model chosen in the end is that of Gradient Boosting which presents the best performance. All the models (to a lesser extent that of Neural Networks) are however successful in their predictions. If the preponderant criterion for choice making is the interpretability then the GLM Gamma model is the optimal one.

The Gradient Boosting model chosen presents an excellent performance on the Backtesting database, i.e. that of the year 2020. This, in addition to the homogeneity of the data, a good performance of simple models (such as linear regression) and good cross-validation results

eliminate the risk of overfitting. This made it possible to validate the model.

This study was carried out on a large number of data. Its results are quite consistent with the theoretical analysis. The predictions are excellent and the chosen model has been validated more than once: on the test database then through Backtesting. The results are therefore reliable.

However, several choices were made in this study, including the variables to be used and the Machine Learning approach. All these choices can be questioned and altering them could result in a better proxy for the BEL. Finally, it would be interesting to extend this study to produce proxies for other indicators such as the SCR, for example.

Glossaire

ALM : Assets Liabilities Management
BEL : Best Estimate Liabilities
PMVL : Plus ou Moins Values Latentes
PPE : Provision Pour Excédents
ESG : Economic Scenario Generator
PB : Participation aux Bénéfices
PDD : Provision pour Dépréciation Durable
PM : Provision Mathématique
PRE : Provision pour Risque d'Exigibilité
RC : Réserve de Capitalisation
SCR : Solvency Capital Requirement
TMG : Taux Minimum Garanti
VA : Valeur Actualisée
MV : Market Value

Table des matières

Remerciements **3**

Résumé **4**

Abstract **5**

Note de Synthèse **6**

Synthesis note **12**

Glossaire **18**

Table des matières **19**

Introduction **23**

1 Le cadre de l'étude **24**

1.1 L'assurance vie 24

1.1.1 L'épargne 24

1.1.2 Les fonds Euro et en Unités de Compte 25

1.2 Le contexte règlementaire 26

1.2.1 Solvabilité II 26

1.2.2 Le BEL en tant qu'indicateur 28

1.3 Le contexte ALM Cardif : deux modèles 32

1.3.0.1 French 34

1.3.0.2 ALS 34

2 La présentation du modèle 36

2.1 Enjeux et difficultés 36

2.1.1 Objectif : vers des calculs plus rapides 36

2.1.2 Solution : un prédicteur d'apprentissage statistique 37

2.2 Etudes des sensiblités au marché à travers différentes variables 37

2.2.1 Les variables 37

2.2.1.1 Les variables primaires dans le modèle Cardif 38

2.2.1.2 Le flexing 47

2.2.1.3 Les variables non choquées 48

2.3 Construction de la base de données 50

2.3.1 Les runs Prophet 51

2.3.2 Les chocs 51

2.3.3 Les limites de la méthode de construction 53

2.3.4 Etude préalable 54

3 L'implémentation du modèle 57

3.1 La modélisation de la courbe des taux : le modèle Nelson Siegel 57

3.1.1 L'utilié du modèle 57

3.1.2 La théorie derrière le modèle 58

3.1.3 L'implémentation du modèle et sa pertinence 58

3.2 Retour au BEL : l'analyse visuelle 59

3.2.1 Une première analyse relative aux variables explicatives 60

3.2.1.1 La sensibilité du BEL aux variables principales 60
3.2.1.2 L'interaction entre les variables 63
3.2.2 Une analyse du BEL de la base de données 64
3.3 Automatisation du proxy 66
3.3.1 Les modèles 66
3.3.1.1 Le Prétraitement des données 67
3.3.1.2 La Régression Linéaire 67
3.3.1.3 Les Modèles Linéaires Généralisés ou GLM 69
3.3.1.4 Le Gradient Boosting 72
3.3.2 La performance des modèles 74
3.3.2.1 Les indicateurs de performance 74
3.3.2.2 Le modèle choisi 76
3.4 La validation du modèle 78
3.4.1 Résultats et discussions 80

Conclusion **81**

Bibliographie **83**

Annexes **85**

A Prérequis de Finance **86**

A.1 Les taux d'intérêt 86
A.1.1 L'actualisation 86
A.1.2 La fréquence de composition 87
A.1.3 La composition continue 87
A.2 Les obligations 88

A.2.1 La valorisation d'une obligation 88

A.2.2 Le coupon couru 88

A.2.3 Les taux zéro-coupon 89

A.2.4 Le calcul des taux zéro-coupon 89

B Prérequis de Statistique **91**

B.1 Test de Kolmogorov 91

B.2 Modèles de Machine Learning basés sur les arbres de décision 92

B.2.1 Decision Trees 92

B.2.2 Adaboost 92

B.2.3 Random Forest 92

C Code Python **94**

D Code R **109**

Introduction

La crise des subprimes qui a eu lieu en 2008 a démontré que la gestion du risque systémique était cruciale. Il s'agit du risque qu'une crise générale se déclenche suite à un enchaînement d'effets négatifs découlant à l'origine d'une faillite d'un membre du système.

Dans la continuité de Solvabilité I (2002) et Bâle II (2007), légiférer pour contrôler les risques pris en terme de fonds propres s'est imposé au centre des débats de réglementation. Parmi les réformes figure la création de l'ACP ou Autorité de Contrôle Prudentiel, prédécesseur de l'ACPR, le régulateur français.

Solvabilité II exige à travers l'un de ses trois volets le calcul d'indicateurs quantitatifs permettant de juger du risque pris pas l'assureur. Un de ces indicateurs est le BEL qui équivaut aux engagements probables de ce dernier. BNP Paribas Cardif a élaboré un modèle appelé ALM visant à projeter de manière stochastique son bilan prudentiel sur un horizon de 40 ans selon les différents scénarios possibles du marché. Cette projection produit en output plusieurs indicateurs dont le BEL, le SCR ou la PVFP.

Le modèle ALM est assez sophistiqué et obtenir des valeurs pour les indicateurs ci-dessus régulièrement est une tâche ardue. La solution apportée à cette contrainte pour les contrats d'épargne en fonds Euro est l'objectif central de ce mémoire. Dans une perspective de rester cohérent par rapport au modèle ALM, la question primordiale est de mesurer la sensibilité du BEL par rapport à la déviation du marché.

La première étape de cette étude consiste en la détermination des variables captant au plus les variations du marché, de les situer dans le modèle afin d'expliciter leurs effets et leurs interactions entre elles et avec le BEL. Le but est ensuite de les dévier de leurs valeurs réelles et d'analyser les conséquences de ceci sur le BEL.

La deuxième étape majeure de ce mémoire établit une base de données à partir de toutes ces valeurs et calibre de multiples algorithmes de Machine Learning sur les données afin d'extraire le modèle optimal permettant d'approximer le BEL.

Chapitre 1

Le cadre de l'étude : le contexte contractuel et règlementaire de Cardif

1.1 L'assurance vie

Le cadre d'étude de ce mémoire est l'assurance vie et plus particulièrement les contrats d'épargne Euro.

En assurance vie il existe deux types principaux de contrats :

— Le premier est un contrat d'épargne permettant le versement d'une somme sous forme de rente ou de capital au bénéficiaire à partir d'une date préfixée dans le contrat tant que l'assuré est en vie ;

— Le second est un contrat de transmission de capital à des bénéficiaires sous forme de rente ou de capital au décès de l'assuré pourvu que cet évènement ait eu lieu après l'arrivée à terme du contrat.

Les contrats d'assurance vie sont un produit très communément souscrit. Présentant des avantages fiscaux, il est utilisé à diverses fins.

1.1.1 L'épargne

Le cadre d'étude de ce mémoire est le contrat d'épargne. Il s'agit d'un type de contrat d'assurance vie qui représente un moyen d'épargner et donc de percevoir une retraite supplémentaire, de se constituer un capital et de le faire fructifier.

D'autres contrats d'assurance vie peuvent présenter d'autres objectids tels que la volonté de protéger des proches (enfants ou ascendants par exemple) contre d'éventuels accidents survenant aux personnes les prenant à charge habituellement.

1.1.2 Les fonds Euro et en Unités de Compte

Les contrats d'assurance vie souscrits sont soit des contrats dits *monosupports* soit *multi-supports*. L'épargne du souscripteur est alors respectivement investie dans un fonds unique ou plusieurs fonds différents appelés supports ce qui permet entre autres de diversifier le portefeuille de l'assuré. Les deux types de fonds sont le *fonds Euro* et le fonds dit en *Unités de Compte* ou *UC*.

Fonds Euro :

Ces contrats se nomment ainsi car la garantie qu'offre l'assureur au bénéficaire s'exprime en euros. Il s'agit d'un fonds moins risqué que celui en *UC*. Concrètement, l'organisme d'assurance est tenu de revaloriser le capital du souscripteur chaque année au *TMG* ou *Taux Minimum Garanti* (détaillé dans la section 2.2.1.1) augmenté de la participation aux bénéfices (voir section 2.2.1.1). L'investissement se fait principalement dans des produits financiers considérés comme étant à risque faible (tels que des obligations d'Etat ou des obligations françaises peu risquées). Le risque est alors complètement absorbé par l'assureur. C'est le fonds d'étude de ce mémoire.

Fonds en Unités de Compte :

Ces contrats se nomment ainsi car la garantie qu'offre l'assureur au bénéficaire s'exprime en unités de compte de valeurs cotées sur le marché, indépendamment de la valeur effective de ces unités. Il s'agit d'un fonds plus risqué que celui en *Euro*. Concrètement, l'organisme d'assurance investit un nombre d'unités dans divers supports financiers. L'investissement se fait principalement dans des produits financiers considérés comme étant plus risqués (tels que des obligations, des actions, des OPCVM [(1)] ou des FCP [(2)]). L'assuré subit et les plus-values et les moins-values de son portefeuille. Le risque est alors complètement absorbé par l'assuré. Afin de protéger l'assuré, des clauses optionelles existent comme la *garantie plancher* qui peut être souscrite afin de sécuriser au moins le remboursement des primes si l'assuré décède.

(1). Organismes de placements collectifs en valeurs mobilières.
(2). Fonds Commun de Placement.

1.2 Le contexte règlementaire

Solvabilité II ou Directive 2009/138/CE est pour le monde de l'assurance que qu'est Bâle II pour le monde bancaire. Cette directive allie exigences quantitatives, qualitatives et de reporting afin de contrôler les risques liés aux activités d'assurance et de réassurance. La mise en place de celle-ci a été fortement suscitée par le contexte de crise de 2008 soulignant l'importance de la prévention contre le risque systémique [(3)]. Ce contexte marque notamment la création de l'*ACP* ou *Autorité de Contrôle Prudentiel* (prédécesseur de l'*ACPR* citée dans la section 1.2.1).

1.2.1 Solvabilité II

Solvabilité II est une règlementation européenne. Elle est entrée en vigueur le 1er janvier 2016. Succédant à et dans la continuité de Solvabilité I, elle vise aussi à assurer la solvabilité, c'est-à-dire la capacité à faire face aux engagements, des compagnies d'assurance et de réassurance. Elle s'inscrit donc dans un objectif de protection des assurés mais aussi d'harmonisation des normes européennes.

Afin de respecter Solvabilité II, une compagnie doit posséder un capital suffisant (reflété pratiquement par ses fonds propres) par rapport aux risques qu'elle encourt. Concrètement, Solvabilité II se base sur trois piliers qui doivent être implémentés.

Pilier I, le pilier quantitatif :

Il définit différents indicateurs calculables permettant de quantifier l'exposition au risque de l'entreprise à travers ses engagements et ses atouts. Concrètement, le pilier détaille le calcul des exigences en fonds propres qui consitutent le capital de l'entreprise ou la somme des apports des actionnaires et des gains générés par l'activité. Ces exigences sont exprimées entre autres à travers deux indicateurs principaux : le *SCR* et le *MCR*.

Le *MCR* ou *Minimum Capital Requirement* définit un montant de fonds propres que doit posséder l'entreprise en dessous duquel il est inenvisageable que l'entreprise poursuive son activité car le risque encouru est trop considérable.

Le *SCR* ou *Solvency Capital Requirement* est égal au capital requis pour faire face à tous les engagements sur un an avec une probabilité de 99.5%, c'est-à-dire au capital nécessaire pour réduire la probabilité de faillite à 1/200. Ce dernier peut se calculer selon deux formules : la formule standard totalement explicitée par Solvabilité II ou le modèle interne qui peut être

(3). Risque qu'une crise générale se déclenche suite à un enchaînement d'effets négatifs découlant à l'origine d'une faillite d'un membre du système.

personnalisé par l'entreprise selon ses besoins sous réserve de validation du modèle.

Il est aussi indispensable de calculer les provisions techniques dans le cadre de ce pilier dont l'indicateur quantitatif principal pour ce mémoire : le *BEL*. Solvabilité II impose les règles de calcul de ces derniers par exemple à travers une valorisation dite « Market consistent » c'est-à-dire basée sur les informations fournies par le marché financier. C'est un point de différence avec Solvabilité I.

Pilier II, le pilier qualitatif :

Il définit l'organisation interne de l'entreprise afin d'implémenter Solvabilité II et de contrôler la qualité de cette implémentation. La démarche *ERM* ou *Enterprise Risk Management* est encouragée dans ce cadre. Elle a pour objectif d'identifier les différents risques de l'entreprise, leurs origines, leurs éventuelles conséquences et de fixer l'appétence au risque.

L'*ORSA* (*Own Risk and Solvency Assessment*) fait partie de cette démarche *ERM* et recalibre les formules proposées par rapport aux risques propres à l'entreprise à travers par exemple une modification de la matrice de corrélation de ces derniers ou une création de nouveaux risques.

Dans le cadre de ce même pilier, une mise en place d'un contrôle interne est alors imposée. L'entreprise est par ailleurs contrôlée par un superviseur externe (*l'Autorité de Contrôle Prudentiel et de Résolution* ou *ACPR* en France) à travers un processus harmonisé par Solvabilité II. Ce processus vérifie la qualité des données, la précision des calculs et la fiabilité des estimations. Cette autorité peut impose un « Add-On » ou surplus de capital si elle juge que l'entreprise a sous-estimé ses risques.

Pilier III, le reporting :

Il permet de répertorier ce processus et d'harmoniser les diverses façons de le faire afin de pouvoir reporter à l'autorité de contrôle. Ce pilier permet aussi une certaine transparence dans le marché à travers des publications de divers indicateurs Solvabilité II reflétant la situation de l'entreprise face à ses concurrents.

L'étude menée dans ce mémoire se concentre sur le premier pilier et plus précisément sur le BEL. Ce dernier fait partie des provisions techniques du bilan de l'entreprise comme schématisé au niveau de la figure 1.1.

Les provisions techniques constitutent une réserve financière qui servirait afin de faire face aux engagements de l'assureur envers les assurés. Ces dernières doivent être estimées afin de calculer le Passif et ainsi le SCR, indispensable en Solvbailite II. Il existe une multitude de types de provisions techniques en assurance vie dont la Provision Mathématique, la Provision

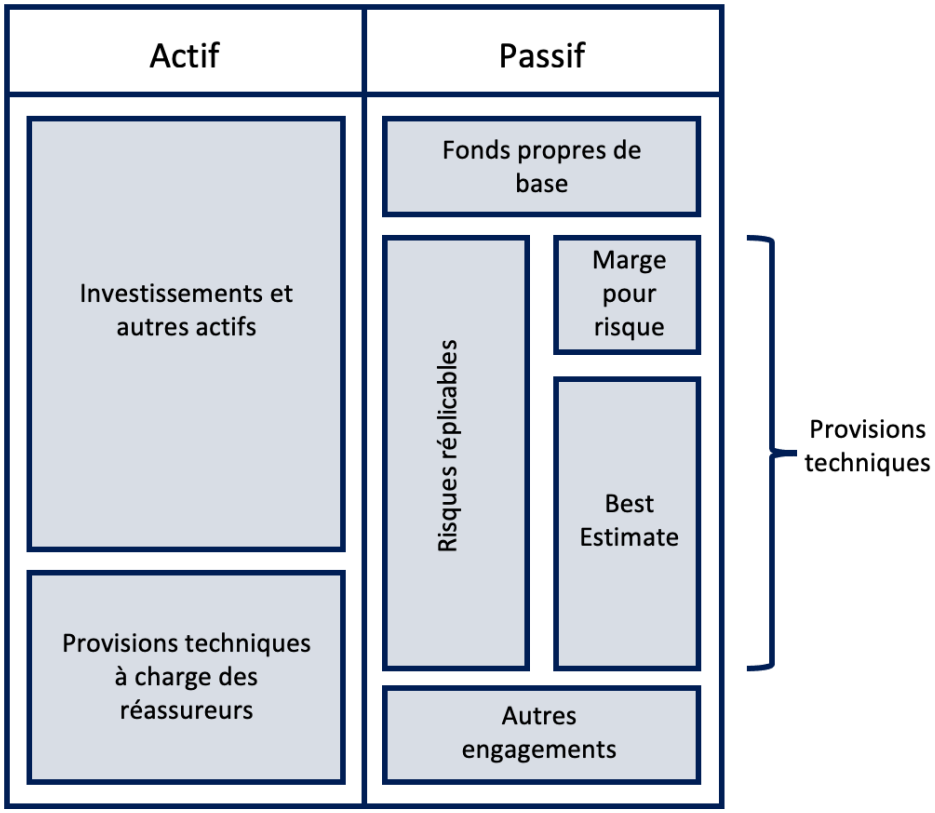

FIGURE 1.1 – Le bilan tel que défini par Solvabilité II

pour Participation aux Excédents, la Réserve de Capitalisation, la Provision pour Dépréciation Durable, la Provision pour Risque d'Exigibilité (voir 2.2.1.2), la Provision pour Aléas Financiers ou encore la Provision Globale de Gestion.

1.2.2 Le BEL en tant qu'indicateur

Dans le cadre du calcul des provisions techniques requis par Solvabilité II, il est indispensable, comme indiqué précédemment, de calculer le BEL ou Best Estimate Liabilities. Cet indicateur est défini par la Directive Solvabilité II qui est ensuite transposée au Code des Assurances français. Ce dernier définit le calcul du BEL à travers l'article R351-2 comme : « la moyenne pondérée par leur probabilité des flux de trésorerie futurs compte tenu de la valeur temporelle de l'argent estimée sur la base de la courbe des taux sans risque pertinente, soit la valeur actuelle attendue des flux de trésorerie futurs ».

Plus simplement, le BEL représente les flux éventuels pondérés par leur probabilité d'occurence et actualisés selon le taux sans risque. Les principaux flux entrant en jeu dans le calcul

du BEL sont les engagements futurs envers les assurés sous forme de prestations, les primes à venir et les coûts futurs. Une augmentation du BEL reflète un grand nombre de flux sortant donc.

Un modèle de projection fournit des captures des portefeuilles de la compagnie (des actifs ou des contrats) en respectant les contraintes de normes auxquelles reste soumis l'assureur sur une durée de projection déterminée. Il permet à l'assureur d'avoir un aperçu de l'avenir afin de mieux se préparer au jour d'aujourd'hui. Le modèle de projection de BNP Paribas Cardif effectue des projections sur 40 ans des flux cités ci-dessus entre autres. Ainsi, une formule fermée du BEL au temps t serait :

$$BEL_t = Valeur\ actuelle(Prestations\ futures) + Valeur\ actuelle(Coûts\ futurs)$$

$$-Valeur\ actuelle(Primes\ futures)$$

Le BEL représente plus généralement les engagements de l'assureur. C'est un indicateur assez parlant d'où sa centralité.

Les prestations

Les prestations sont dûes aux assurés en cas de matérialisation d'un sinistre. Dans le cas des contrats d'épargne, ces dernières sont surtout souvent liées au décès de l'assuré ou à sa survie, à l'arrivée à maturité d'un contrat et à la possibilité de rachat de la totalité ou d'une partie du contrat.

Le calcul des prestations envers les assurés se calcule plus précisément comme suit :

$$Prestations\ futures = Décès + Rachats\ structurels + Rachats\ dynamiques$$

$$+Maturités + Arbitrages + Transferts + Prestations\ de\ fin\ de\ projection$$

Le terme Décès englobe les montants qui doivent être versés aux bénéficiaires suite au décès de l'assuré.

Les rachats structurels ou statiques sont les prestations dues aux assurés ayant choisi de retirer une partie ou la totalité de l'épargne versée en vue d'assurer un besoin de liquidité et ce avant le terme du contrat. Les rachats structurels sont soit partiels (le contrat prévoit une possibilité pour l'assuré de retirer une partie de son épargne), totaux (l'assuré peur retirer toute l'épargne) ou réguliers (l'assuré peut retirer régulièrement une partie de l'épargne selon un taux

de rachat bien défini). Ce terme est estimé dans le modèle *French* (voir 1.3.0.1) de BNP Paribas entre autres à travers les tables de mortalité.

Les rachats dynamiques sont les rachats dus à des répercussions d'aléas économiques. Ils sont liés à l'écart entre le taux servi concrètement (qui dépend du processus de Participation aux Bénéfices expliqué) et le taux cible compétitif. Ils peuvent être partiels ou totaux. Ce terme est estimé dans le modèle *ALS* (voir 1.3.0.2) de BNP Paribas entre autres à travers la table Economic Scenario Generator (voir 2.2.1.1).

Le terme Maturités comprend les prestations que l'assureur doit verser suite à l'arrivée à maturité des contrats. Ces prestations peuvent être versées en une fois sous forme de capital ou en plusieurs prestations appelées rentes.

La partie Arbitrages sont des frais éventuels découlant d'un transfert de la totalité ou d'une partie du capital de et vers un autre fonds (du fonds Euro en unités de compte et vice versa).

Les transferts sont la perte engendrée lorsque des assurés souhaitent transférer leur capital à un autre assureur.

Le modèle BNP Cardif ne projettant que sur 40 ans, les gains restant au terme de cete projection sont partagés avec les assurés sous forme de prestations de fin de projection. Ces prestations se déclinent sous plusieurs types dont la Provision Mathématique (PM), la Réserve de Capitalisation (RC), la Provision pour Dépréciation Durable (PDD), la Provision pour Risque d'Exigibilité (PRE) ou encore la Plus ou Moins Value Latente (PMVL).

Les primes

Les primes sont un flux positif pour l'assureur versé par le souscripteur du contrat. Ce montant peut être versé de manière régulière (versement périodique mensuel, trimestriel ou annuel), libre (selon la décision préalable du souscripteur) ou unique (un seul versement).

Les coûts

Plusieurs coûts sont à charge de l'assureur tels que les commissions, les frais généraux, les taxes et les coûts de gestion des contrats.

Afin de cerner plus en profondeur le calcul du BEL, il est nécessaire de distinguer les deux profils de risques principaux. Pratiquement, il est possible de distinguer deux catégories de risques : les risques couvrables (dits *headgeable*) et les risques non couvrables. Les provisions citées ci-dessus répondent à une volonté de se prémunir de diverses combinaisons de ces deux risques.

Les risques couvrables

Les risques couvrables correspondent aux risques vus comme modélisables et/ou prévisibles. Il s'agit du risque résultant d'aléas économiques. Souvent l'assureur y est exposé à travers des instruments financiers. Se couvrir contre ces risques revient à élaborer des stratégies financières statiques ou dynamiques [4] afin de mimer [5] le comportement du marché et de limiter l'écart [6] par rapport à ce dernier. Evaluer ces risques s'effectue pratiquement grâce aux calcul de l'espérance des ces flux actualisés sous la probabilité dite risque neutre [7]. La provision récapitulant ces risques est le BEL.

Les risques non couvrables

Les risques non couvrables correspondent aux risques vus comme imprévisibles et/ou non modélisables. Il s'agit de risques résultant d'évènements divers. Il s'agit des risques de rachats statiques (liés aux besoins des assurés que l'assureur ignore a priori), de mortalité, de contrepartie. Se couvrir contre ces risques revient à constituer une marge afin de les absorber. C'est la *Marge pour Risque*.

Les provisions techniques du Passif du bilan Solvabilité II de l'assureur se compose donc de la somme de ces deux marges : le BEL et la *Marge pour Risque* comme schématisé dans la figure 1.1. La *Marge pour Risque* n'est pas l'objet d'étude de ce mémoire qui s'intéresse exclusivement au BEL.

Ainsi, le BEL est calculé à travers la quantification des prestations, coûts et primes. Cependant, cette quantification ne doit prendre en compte que les provisions des risques couvrables. Par ailleurs, le calcul s'avère complexe à travers une formule fermée. En effet, les prestations par exemple dépendent des rachats dynamiques qui sont les réalisations de variables aléatoires à modéliser selon les scénarios économiques et les tables de mortalité par exemple.

Afin de répondre à cette problématique, le BEL est concrètement calculé autement notamment à travers une approche Monte Carlo et non une formule fermée selon l'expression suivante du BEL au temps t :

$$BEL_t = \mathbb{E}\left(\sum_{i=0}^{40} F_i.a_i\right)$$

où :

(4). Une stratégie statique, contrairement à une stratégie dite dynamique, n'est pas ajustée au fur et à mesure de l'évolution du marché.

(5). Le terme financier est la réplication d'un produit financier.

(6). Cet écart est communément appelé PnL ou Profit & Loss en Finance.

(7). Le calcul de l'espérance se fait plus facilement grâce à un changement de probabilité vers une probabiltié dite risque neutre.

— F_i est l'ensemble des flux de l'année i pondérés par leurs probabilités d'occurence. Les flux de l'année i étant les primes, les prestations et les dépenses de l'année en question ;

— a_i est l'actualisation selon le taux sans risque de l'année i.

Dans la suite, le BEL est principalement considéré comme étant représentatif des engagements de l'assureur puisque ceux-ci composent une partie considérable du BEL.

1.3 Le contexte ALM Cardif : deux modèles

Afin de s'intéresser de plus près au calcul du BEL, il est nécessaire de se focaliser sur l'*ALM* ou *Asset Liabilities Management*. Comme son nom l'indique, c'est un processus qui vise à modéliser le Passif et l'Actif. Pratiquement, et afin de produire une multitude d'indicateurs tels que le BEL et le SCR, l'ALM de Cardif projette le Passif et l'Actif en prenant en compte l'interaction entre les deux et ce à travers le logiciel Prophet.

Cette projection se fait sur un certain horizon prédéterminé (40 ans chez Cardif) sur lequel se fait la prédiction de l'évolution du Passif et du portefeuille d'Actif en prenant en compte plusieurs facteurs qui pourraient influencer cette évolution. Ceci se fait à travers deux modèles explicités ci-après qui allient projections déterministe et stochastique.

L'ALM de Cardif peut être résumé par la figure 1.2. Il se décompose en deux sous-modèles majeurs : *French* et *ALS*. *French* modélise le Passif en prenant en compte différents inputs et fournit des outputs qui sont pris en input dans le modèle *ALS*. Ce dernier modélise l'Actif tout en prenant en compte l'interaction Actif-Passif. Il fournit plusieurs indicateurs en output nécessaires à la constitution du bilan de l'assureur.

L'ALM permet de répondre donc aux exigences de Solvabilité II en fournissant divers indicateurs dont le SCR et le BEL mais aussi d'aiguiser sa stratégie d'investissement et d'optimiser l'allocation d'actifs (plus de détails ici 2.2.1.1). En effet, il offre une vision claire du bilan en le projetant selon divers scénarios économiques pour chaque composition spécifique en actifs du portefeuille.

Le modèle de projection ALM de Cardif est GPM ou Group Projection Model. Il est implémenté comme décrit précédemment sur Prophet. Ce dernier est l'outil qui permet d'implémenter ces modèles, d'effectuer des calculs et des projections qu'elles soient stochastiques (méthode de Monte Carlo) ou déterministes selon s'il reçoit en input un seul ou plusieurs jeux de données. Il est particulièrement utile en provisionnement et tarification.

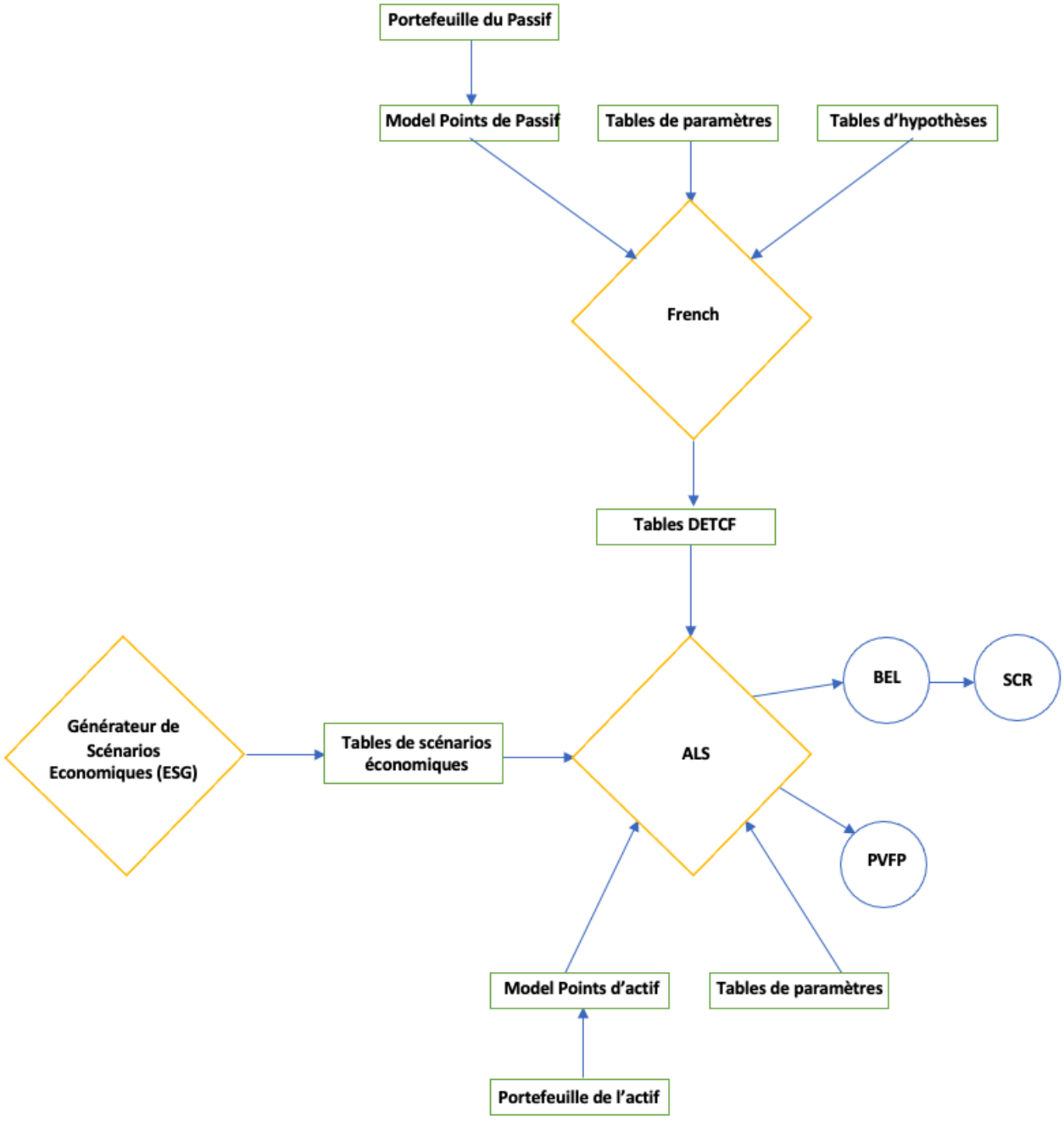

FIGURE 1.2 – Le schéma ALM de BNP Paribas Cardif

1.3.0.1 French

French est le modèle du Passif de BNP Paribas Cardif. Il s'agit d'un modèle déterministe qui fait abstraction des scénarios économiques. Il produit des projections à maille contractuelle et mensuelle.

Les contrats qui constituent le portefeuille du Passif sont regroupés dans des groupes homogènes à une faible granularité selon une multitude de caractéristiques en commun (TMG et âge des assurés par exemple) qu'on nomme *poches stochastiques*. Ce sont les *Model Points du Passif*. Ils constituent un des inputs du modèle *French*.

Le modèle *French* prend ensuite en input des tables d'hypothèses et des tables de paramètres contenant des tables de mortalité, des lois de rachat, les provisions mathématiques en début de projection (voir 2.2.1.3) et calcule en output une projection des flux du Passif tels que des rachats structurels, des frais et autres flux négatifs. Cette prjection est déterministe. Elle ne prend pas en compte les aléas économiques mais elle se base plutôt sur les caractéristiques des contrats souscrits (âge des assurés, montant de la prime et fréquence par exemple).

Les flux du Passif en output sont agrégés à une maille annuelle par souci d'épargne de temps de calcul dans des tables appelées tables *DETCF* (voir figure 1.2). Il est important de noter de même que, la Provision Mathématique étant l'engagement de l'assureur envers l'assuré, le modèle en question permet de calculer entre autres ce montant projeté au TMG (Taux Minimal Garanti).

1.3.0.2 ALS

Le modèle *ALS* est le modèle Actif-Passif de BNP Paribas Cardif. Il s'agit d'un modèle stochastique prenant en compte les scénarios économiques et les tables *DETCF* du modèle *French* à une maille annuelle afin de produire divers indicateurs dont le BEL.

Les divers actifs du groupe qui constituent le portefeuille d'actifs sont regroupés dans des groupes homogènes selon le type d'actif auquel ils appartiennent. Ce sont les *Model Points d'Actif*. Parmi ces derniers figurent par exemple les *Model Points* d'obligations, d'actions ou de produits dérivés. Ils constituent un des inputs du modèle *ALS*. Le modèle Cardif distingue 13 types d'actifs. Les tables détaillent la valeur comptable de l'actif (valeur effective du bien dans le bilan), sa valeur de marché (valeur qu'il a aujourd'hui sur le marché) mais aussi d'autres caractéristiques telles que la distribution ou non d'éventuels coupons ou dividendes.

Les *Model Points* d'obligations regroupent notamment les obligations gouvernementales,

corporate, à taux fixe et à taux variable. Les *Model Points* d'actions sont constitués principalement d'actions rachetables et non rachetables, de sociétés cotées. Quant aux produits dérivés, il s'agit par exemple de caps et de floors.

Le modèle *ALS* prend ensuite en input des tables de paramètres et les tables ESG qui énumèrent une multitude de scénarios. *ALS* tient compte de ces aléas économiques afin de calculer divers indicateurs pour chaque scénario, qui sont souvent calculés les uns en fonction des autres et modélise dans le processus les rachats dynamiques. Pratiquement, pour le BEL, Cardif doit calculer 152 indicateurs pour le fonds Euro afin de calculer le BEL et ce pour chaque scénario. Etant donné que Cardif utilise 1000 simulations pour la méthode Monte Carlo, ceci équivaut à 152 000 indicateurs au total pour calculer le BEL. Dans le but de produire un résultat plus agrégé, il est souvent plus pratique de conserver en tête surtout l'équivant certain ou scénario *EC* qui consiste à conserver la moyenne des 1000 simulations.

ALS permet aussi de modéliser la gestion de l'assureur au niveau des actifs et ce à travers la stratégie financière (voir 2.2.1.1), processus qui vise à imiter l'action des gestionnaires d'actifs.

Globalement, *ALS* projette en résultat et de manière stochastique l'Actif et le Passif en prenant toutes les données citées. C'est l'output de ce modèle qui fera donc surtout l'objet de notre étude.

Chapitre 2

Les objectifs du modèle, les variables choisies et l'étude préalable

2.1 Enjeux et difficultés

Produire divers indicateurs pour Solvabilité II est une obligation pour les assureurs. Cependant, ces calculs sont très coûteux en temps et les modèles utilisés tels ceux décrits précédemment sont souvent complexes. Ce mémoire vise à résoudre une partie de ce problème.

2.1.1 Objectif : vers des calculs plus rapides

Il est utile de produire des résultats sur une maille de temps plus petite. En effet, ceci permet de bénéficier d'une marge temporelle plus grande pour changer sa stratégie actuelle. Il offre à la direction une meilleure trame pour effectuer des décisions plus ciblées et basées sur des données concrètes et quantifiables. Par ailleurs, ceci crée la possibilité de prédire les conséquences de divers scénarios probables sur les indicateurs importants. C'est dans ce cadre général que s'inscrit l'utilité d'un proxy (ou approximation), outil qui permet d'estimer ces indicateurs et ainsi de mieux gérer ses risques. Dans ce mémoire, le proxy est effectué sur le BEL. Toutefois, ce travail précède plusieurs futurs travaux de proxy (du SCR notamment) que compte mener Cardif.

2.1.2 Solution : un prédicteur d'apprentissage statistique

La solution qui a été pensée pour ce problème est l'implémentation d'un proxy de Machine Learning. Ce travail suit l'esprit d'un mémoire réalisé auparavant par Meriem Araqi Houssaini intitulé « Anticipation de la déviation du BEL suivant différents états du monde »[(1)]. Des variables ayant un poids considérable sur la déviation du BEL sont choisies puis choquées afin de constituer une base de données qui traduit la déviation de cet indicateur selon la variation de ces variables. Cette base de données est ensuite analysée et soumise à divers algorithmes d'apprentissage statistique afin de déceler les liens potentiels entre les variables explicatives et la variable à prédire.

L'approche de type stress test ou choc est classique en actuariat puisqu'elle permet par exemple de calculer le SCR. Elle permet de prédire les déformations de l'indicateur en question et de projeter le bilan dans une éventualité choisie. Ces chocs modélisent les fluctuations des variables clés suite à des phénomènes financiers, économiques ou sociaux.

2.2 Etudes des sensiblités au marché à travers différentes variables

La portée d'étude de ce mémoire est celle de l'effet des aléas du marché sur le BEL. Comme expliqué précédemment dans la section 1.2.2, cerner le BEL revient à cerner principalement les risques couvrables ce qui explique l'intérêt pour les aléas du marché. Dans ce qui suit, ce mémoire vise à spécifier un plan d'action de la solution proposée ci-dessus. On se penche ici sur les variables qui reflètent au mieux le risque marché mais qui semblent aussi englober divers aspects de ce dernier et traduire au mieux sa complexité. Ces variables doivent aussi être prépondérantes dans le calcul du BEL. On essaie de les situer ensuite dans le modèle afin de comprendre leur calcul, ce qu'elles représentent et leur poids sur cet indicateur. Leur analyse permettra de prédire les résultats auxquels on s'attend avant de mener l'étude quantitative.

2.2.1 Les variables

Les trois premières variables abordées sont les variables centrales de cette étude, à savoir la PMVL, la PPE et la courbe de taux zéro-coupon. Ces variables sont celles qui ont été choisies afin de représenter le marché. En réalité, elles reflètent respectivement l'effet de la volatilité de

(1). HOUSSEINI (2016).

la valeur de marché des actions, celui de la participation aux bénéfices et celui de la variation des taux sur les obligations (qui représentent en général 80% du portefeuille de Cardif). Ce sont ces variables qui seront sujettes à des chocs et qui alimenteront principalement la base de données constituée par la suite.

On s'est penché aussi sur d'autres variables telles que la PM, la PDD, la RC et le PRE. Ces dernières n'ont pas fait l'objet de la même étude que celle menée sur la PMVL, la PPE et la courbe de taux (voir 2.3) car elles traduisent moins directement l'effet de marché sur le BEL.

2.2.1.1 Les variables primaires dans le modèle Cardif

Les variables dont il est question dans cette section sont la PMVL, la PPE et la courbe des taux zéro-coupon. Afin de comprendre le rôle de ces variables dans le calcul du BEL, il est important de s'intéresser à la stratégie financière (figure 2.1). Cette dernière fait partie intégrante du modèle ALS de la compagnie et elle est très essentielle pour notre étude.

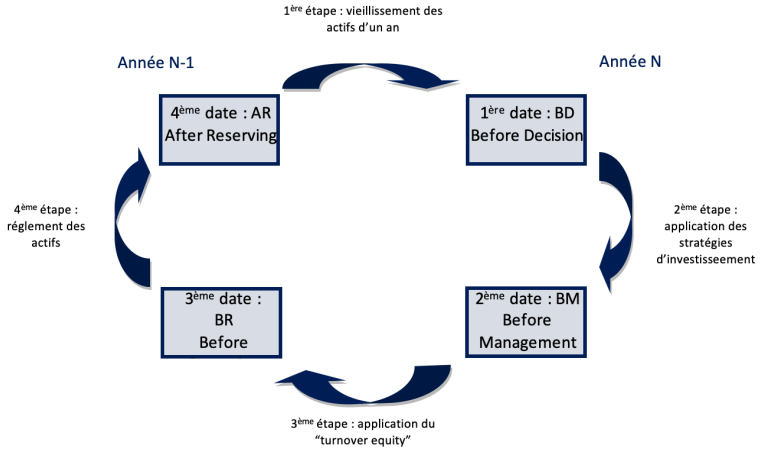

FIGURE 2.1 – Les étapes principales de la stratégie financière

Pendant la projection effectuée dans ALS, un processus nommé stratégie financière est implémenté. Il permet la modélisation de l'actif et plus précisémet du comportement des gestionnaires d'actifs, c'est-à-dire l'allocation d'actifs (la composition du portefeuille d'actifs) et les achats et ventes à effectuer à chaque fin d'année. Le processus permet surtout de vérifier en continu l'adéquation à diverses contraintes auxquelles reste soumis l'assureur (de solvabilité et de distribution des bénéfices notamment). Le respect des exigences de la directive Solvabilité II entraîne plusieurs contraintes sur la gestion d'actifs et vice-versa. En effet, l'assureur peut être amené à gérer d'une certaine manière son portefeuille d'actifs afin de minimiser le SCR. Il faut aussi respecter la composition du portefeuille d'actifs définie par la direction et équilibrer

le bilan en égalisant l'Actif et le Passif, à chaque année de projection du GPM.

C'est ce que permet d'assurer la stratégie financière qui effectue ceci pour chaque scénario économique de la table ESG à travers quatre étapes qui ont lieu à chaque fin d'année. Les variables principales de l'étude sont calculées à travers ces dates dans la stratégie financière. Les quatre dates sont :

BD ou *Before Decision* :

Le portefeuille d'actifs est vielli d'un an au sens de l'actualisation afin de prendre en compte sa vraie valeur actuelle. Aucune décision de gestion (d'achat ou de vente) d'actif n'est encore prise d'où le nom de l'étape.

BM ou *Before Management* :

Certains flux tels que les prestations, les primes et les charges ont lieu. La stratégie d'investissement est ensuite appliquée. Elle détermine la composition du portefeuille en chaque type d'actif. Il peut s'agir de la stratégie fixe (qui impose des proportions pour chaque actif à respecter) ou la stratégie Tunnel (qui impose un intervalle de composition pour chaque actif). Des achats/ventes ont alors lieu afin de se plier à cette ré-allocation d'actif. Le portefeuille d'actifs est donc impacté ici. On est à l'étape qui précède la gestion des plus ou moins values éventuelles.

BR ou *Before Reserving* :

Le turnover a lieu. Il s'agit de l'achat et la vente rapides d'actifs afin de générer des bénéfices immédiats, tout en n'altérant pas la quantité de chaque actif. Des produits financiers, gains et pertes en résultent. Puisqu'il y a achat d'actifs, il y a modification des valeurs comptables de ces derniers. On n'a cependant toujours pas pris en compte le Passif. On est donc en *avant provisionnement*.

AR ou *After Reserving* :

On prend en compte le provisionnement comme le mécanisme de participation aux bénéfices. On connait alors la valeur du Passif ce qui permet le réalignement de l'Actif. Ce dernier consiste à égaliser l'Actif et le Passif afin d'équilibrer le bilan tout en respectant la stratégie d'investissement.

Il est incontestable qu'évaluer le BEL nécessite alors une analyse de la stratégie financière puisqu'elle résume l'interaction Actif-Passif et la mise à jour de ces derniers selon les scénarios économiques.

Il est important de s'intéresser pour le moment à la $1^{\text{ère}}$ étape de la stratégie financière :

celle du vieillissement d'actifs. C'est l'étape qui a lieu lors du passage de la dernière date de l'année N-1 (*After Reserving*) à la première date de l'année N (*Before Decision*).

Le vieillissement d'actifs réévalue les actifs. Selon le type d'actif, il peut s'agir de recalculer la valeur comptable ou la valeur de marché (qui a certainement évolué depuis un an). Ceci dépend principalement du classement de l'actif en amortissable ou non amortissable. Cette distinction est régie par les normes comptables françaises et elle est importante par la suite afin de bien comprendre le rôle des variables dans le processus.

Les actifs amortissables perdent de la valeur naturellement. Il est alors indispensable de prendre en compte cette dépréciation au niveau du bilan. Il peut s'agir de biens qui subissent de l'usure, qui deviennent obsolescents ou qui ont moins de valeur avec le temps. Leur valeur comptable s'écarte donc avec la dépréciation de leur valeur d'acquisition et il faut recalculer la première régulièrement. Les titres amortissables sont principalement les obligations chez un assureur.

Les titres non amortissables sont par exemple les actions. Leur valeur comptable est la même que leur valeur d'acquisition (ceci est théoriquement vrai mais pratiquement à nuancer voir le paragraphe 2.2.1.1 concernant la PMVL). Cependant, ces dernières s'écartent rapidement de leur valeur de marché qu'il faut reprendre en compte régulièrement.

La PMVL

La variable PMVL ou Plus ou Moins Values Latentes intervient dans la 1ère étape du vieillissement d'actifs. Ce sont des *Unrealized Gains* c'est-à-dire des gains (arithmétiques) potentiels. Ils concernent seulement les actions (*Equity*) et représentent les gains ou pertes que pourrait se faire l'assureur s'il vendait ou achetait certaines actions. Ces gains ou pertes restent fictifs et peuvent se matérialiser plus tard dans le processus à travers le turnover (3ème étape). Il est important de modéliser l'effet de la variation de la valeur des actions sur le marché car ces derniers sont assez volatiles.

La PMVL est pratiquement le résultat de l'écart entre la valeur de marché et la valeur comptable (qui est la même que la valeur d'acquisition dans ce cas puisqu'il s'agit d'actions) selon la formule :

$$PMVL_N = MV_N - FAV_N$$

où MV désigne la Market Value (ou valeur de marché) de l'actif pour l'année N et FAV

désigne la Fair Asset Value (ou valeur comptable) de l'actif pour l'année N.

Tout potentiel découle ainsi purement de la variation de la valeur de marché si l'on suppose la valeur comptable constante. En réalité, la valeur de marché et la valeur comptable varient car on réinvestit les dividendes des actions dans ces mêmes actifs. Il s'agit d'un choix de valorisation des actifs effectué par Cardif. Une alternative aurait été de percevoir simplement les dividendes en *Cash*. Pour la valeur comptable donc on a simplement :

$$FAV_N = FAV_{N-1} + Dividendes_N$$

Pour la valeur de marché, cette relation n'est pas valable car d'une année à une autre, en plus du réinvestissement des dividendes, les aléas du marché modifient la valeur de l'action ce qui crée des PMVL. On a donc dans ce cas :

$$VM_N = VM_{N-1} + Dividendes_N + PMVL_N$$

C'est le vieillissement des actions. La PMVL permet de s'affranchir donc de ce terme des dividendes puiqu'il est éliminé.

Afin de simuler les aléas du marché des actions, il suffit donc de choquer la PMVL ce qui revient à choquer la valeur de marché de l'action.

Il est important de noter l'utilité de ce choc dans la globalité du modèle. Il permet de modéliser l'effet sur le BEL d'une modification d'une partie des inputs du modèle ALS : les actions dans les Model Points d'actifs. Des chocs seront effectués plus tard, à travers une autre variable sur les obligations dans les model points d'actifs. Ces deux types d'actifs représentent la majorité du portefeuille de Cardif et les choquer revient à faire le point sur la volatilité de tout le portefeuille d'actif.

La PPE

L'intérêt se pose à présent sur l'effet de la PPE sur le BEL. Les primes d'un contrat d'assurance vie versées par l'assuré sont investies par l'assureur. Il s'agit de l'épargne des assurés dans notre horizon d'étude [(2)]. Les bénéfices réalisés grâce à ce capital devraient donc profiter, en partie du moins, aux assurés. Effectivement, la loi stipule que ce soit le cas pour la majeure partie de ces derniers. C'est la participation aux bénéfices.

(2). On dit qu'il y a revalorisation de l'épargne.

Légalement, dans le cadre de la participation aux bénéfices et au minimum, 85% des bénéfices financiers [3] et 90% des bénéfices techniques [4] doivent être redistribués aux assurés.

L'assureur peut s'engager dans le contrat à distribuer un taux supérieur au taux garanti légalement [5]. L'assureur ne peut alors distribuer en-deçà de ce nouveau seuil. Généralement, il vise un taux assez attractif afin de rester compétitif et donc de réduire les rachats dynamiques. C'est le *Taux cible*. D'un autre côté, il fixe un *Taux seuil* à ne pas dépasser.

Il est nécessaire de noter que l'obligation légale mentionnée ci-dessus est à nuancer légèrement. En effet, la distribution ne doit pas se faire totalement instantanément. Chaque année, l'assureur distribue la Provision Mathématique (voir 2.2.1.3) augmentée de la participation aux bénéfices. L'assureur dispose de huit ans pour redistribuer la totalité de ce qu'il doit au bénéficiaire du contrat.

Entre temps, l'assureur constitue des provisions pour amortir les éventuels risques auxquels il peut faire face. Parmi ces provisions, on peut citer la PPE ou Provision pour Participation aux Excédents [6]. La PPE est le mécanisme principal utilisé pour lisser le *Taux servi* aux assurés et donc limiter les rachats dynamiques.

Le mécanisme de la PPE fait partie d'un mécanisme plus global : celui de la participation aux bénéfices. Cette stratégie se fait pratiquement en six étapes :

Etape 1 :

L'assureur détermine ses gains. Pratiquement, il s'agit de calculer le TRA [7]. Le TRA pour l'année N (noté TRA(N)) est calculé comme suit :

$$TRA(N) = \frac{Produits\ Financiers\ Disponibles(N)}{AMC(N)}$$

où :

— $Produits\ Financiers\ Disponibles(N) = Cash + PMVL + Cashflows + Obligations - Frais - Variation\ RC - Variation\ PDD$ [8] ;

— $AMC(N)$ est l'*Assiette Moyenne Corrigée* des provisions mathématiques (voir 2.2.1.3)

(3). Il s'agit des bénéfices relatifs à l'investissement de l'épargne des assurés.
(4). Ce sont les gains liés à l'activité-même de l'assurance i.e. la différence entre les ressources (primes et produits financiers par exemple) et les dépenses (prestations, frais de gestion, commissions, etc.).
(5). Ce taux garanti légalement est fixé par le Code des Assurances. Il s'agit du *TMG* ou *Taux Minimum Garanti*.
(6). Cette provision est aussi parfois appelée PPB ou Provision pour Participation aux Bénéfices.
(7). C'est le Taux de Rendement des Actifs. Il reflète le taux de gain ou perte du fonds Euro dans l'année.
(8). La RC et la PDD sont introduites ici 2.2.1.3.

relatives à l'épargne.

Etape 2 :

Cette étape permet de calculer la PB contractuelle qui permet de déterminer le montant à servir selon le contrat. Le calcul de la marge de l'assureur qui pourra être utilisée par la suite comme levier s'effectue également dans cette étape.

Etape 3 :

Le calcul du montant cible est exécuté. Il s'agit du produit entre le *Taux cible* et la PM (voir 2.2.1.3).

Etape 4 :

Cette étape est celle d'intérêt principal pour ce mémoire car elle concerne la *PPE*. Les leviers détaillés ci-contre 2.2.1.1 sont utilisés afin de servir le taux cible.

Etape 5 :

L'étape 5 permet de calculer la PPE_L [(9)].

Etape 6 :

L'étape finale permet de calculer le taux servi effectivement pour l'assuré et la marge finale de l'assureur après bénéfices et potentielle utilisation de cette dernière comme levier aux étapes précédentes.

Le lien entre la PPE et le BEL est alors clair puisque, selon le bilan solvabilité II (voir figure 1.1), déterminer les provisions est essentiel pour déterminer le BEL. Il est alors intéressant de s'attarder plus sur le mécanisme de participation aux bénéfices qui influence la PPE.

L'idée générale est la suivante : durant ces 8 années, la PPE est dotée (ou approvisionnée) lorsque l'assureur jouit de bénéfices excédant les prévisions. Au contraire, il puise dans cette dernière si les rendements sont plus faibles que prévu. Cette stratégie permet d'homogénéiser les rendements pour les assurés et de faire face aux éventuels aléas de rendement. Ceci permet de limiter en partie les rachats dynamiques en cas de situation économique défavorable.

(9). La dotation et la reprise de cette provision sont détaillées dans la suite.

Il est nécessaire d'introduire la différence entre la *PPE_C* (*PPE contractuelle*) et la *PPE_L* (*PPE libre*) afin de clarifier le processus davantage. Tous les calculs suivant s'effectuent à la maille de chaque poche stochastique i (voir 1.3.0.1) :

Le premier levier : la PPE contractuelle

Si la participation aux bénéfices de la poche excède la cible de l'assureur alors la PPE_C est dotée sinon elle est reprise. Cette provision est dotée jusqu'à un seuil spécifique au-delà duquel l'assureur dote une autre provision (la PPE_L) et verse un surplus de bénéfices à l'assuré.

A travers ce mécanisme, le *Taux servi* par l'assureur peut dépasser le *Taux cible*. Dans le cas contraire, il fait usage d'un autre levier afin d'atteindre cet objectif : la *PPE libre*.

Le deuxième levier : la PPE libre

Si l'assureur n'atteint pas le montant cible à servir et que toute la PPE_C est épuisée, il reprend la la PPE_L pour les contrats dits privilège d'abord puis pour les contrats non privilège s'il subsiste encore une partie de la PPE_L.

Le dernier levier : la marge de l'assureur

Un dernier levier qu'utilise l'assureur est sa propre marge de gain au cas où le taux cible n'est toujours pas atteint au terme de l'utilisation des deux précédents leviers.

La variable explicative conservée du mécanisme de la participation aux bénéfices est la PPE_C. C'est la variable qui représente la richesse initiale de l'assureur (ses gains/bénéfices). La PPE_L ne fera plus l'objet de l'étude car elle est négligeable puisqu'elle n'est créée que dans des cas particuliers.

La courbe des taux

Afin de comprendre le rôle de la courbe des taux dans la projection du BEL, il est nécessaire de se pencher sur la table ESG. Cette dernière permet de projeter divers indices financiers (variables économiques et actifs) de manière stochastique en générant divers scénarios possibles d'évolution à travers des modèles mathématiques. Cette table génère notamment des scénarios d'évolution de l'inflation, des rendements des actions, de l'immobilier mais surtout des taux d'intérêts zéro-coupon (voir A.2.3) qui sont notre variable d'intérêt dans cette section. Ce sont des facteurs de risque importants en assurance et qui affectent logiquement le bilan de l'assureur [(10)].

(10). Le BEL étant calculé à partir du bilan de l'assureur, un choc de la table ESG affecte considérablement cette variable.

Par ailleurs, il est essentiel de noter que l'outil qu'est la table ESG permet de projeter les variables en question de manière cohérente avec le marché [(11)].

Etablir une table ESG s'effectue en pratique à travers les cinq étapes principales suivantes :

— Choisir le modèle de projection à appliquer à la variable d'intérêt ;

— Choisir les inputs pertinents pour chaque modèle ;

— Calibrer les modèles ;

— Projeter et générer tous les scénarios stochastiques ;

— Valider les scénarios générés.

A titre d'exemple et afin d'illustrer la complexité du choc de cette table, on cite une multitude de modèles utilisés selon l'indice d'intérêt : souvent un modèle lognormal Black & Scholes pour les actions, une approche Monte Carlo pour les swaptions, une formule fermée pour les Caps et un modèle Cox Ingersoll Ross à 2 facteurs pour les taux. Choquer une table ESG doit tenir compte des interactions entre ces modèles et les variables financières qui la composent, sans oublier de bien fournir la multitude de scénarios qui résultent de ce choc.

Concrètement pour le taux spot [(12)], Cardif choisit de le modéliser à travers un modèle Cox Ingersoll Ross à 2 facteurs ou CIR2++. Ce modèle n'est pas étudié dans ce mémoire.

Notons à présent l'utilité du choc de la courbe des taux zéro-coupon plus précisément dans la globalité du modèle. La sensibilité de ce taux permet de modéliser l'effet sur le BEL d'une modification de deux éléments inputs du modèle ALS : la valeur de marché des obligations (et donc leur valorisation selon A.2 et A.2.3) dans les Model Points d'actifs et plus globalement la table ESG en elle-même (et pour laquelle choquer les taux zéro-coupon se répercute sur toute la table). La figure 1.2 permet de clarifier quels inputs sont impactés dans le processus. Ce qui suit s'intéressera aux détails de ces impacts.

Afin de mieux situer la valorisation des obligations dans le processus ALS, on revient à la 1ère étape du vieillissement d'actifs. On s'intéresse ici au vieillissement des actifs de type obligations (ou *Bonds*). Ces derniers, contrairement aux actions, sont des actifs amortissables (voir 2.2.1.1 pour l'explication des amortissements). Ceci veut pratiquement dire que la valeur comptable évolue d'une année à une autre et s'écarte de la valeur d'acquisition.

La valeur de marché des obligations évolue aussi d'une année à une autre et elle est égale à l'actualisation des flux à venir selon la méthode expliquée dans l'annexe A.2. Il est important de

(11). On dit que la projection est *Market Consistent*.
(12). Il s'agit d'une appellation alternative au taux zéro-coupon.

rappeler que cette actualisation se fait au taux zéro-coupon (A.2.3) qui est la variable choquée ici. Par suite, les conséquences de ce choc sont directes sur la valeur de marché des obligations.

La valeur comptable, quant à elle, est sujette à un amortissement chaque année. Il existe deux types d'amortissement possibles parmi lesquels l'assureur peut choisir afin de modéliser le monde réel : l'amortissement actuariel et l'amortissement linéaire. Le deuxième est plus simple à implémenter mais reste, contrairement au premier, moins indicatif. En effet, l'amortissement actuariel permet d'égaliser pour chaque année la valeur nette comptable et la valeur actuelle des flux futurs au taux de rendement actuariel de l'achat. C'est comme s'il permettait de valoriser l'obligation au même taux d'achat au fur et à mesure du temps en ne prenant en compte que les flux restants. Pour détailler les deux méthodes, on suppose que l'obligation est achetée au temps 0 [(13)]. On a :

L'amortissement actuariel

Avant de pouvoir déterminer l'amortissement actuariel, il est nécessaire de calculer le taux de rendement actuariel de l'obligation. Il s'agit du taux qui permet d'égaliser le prix d'acquisition au temps 0 et la valeur actuelle des flux suivant ce temps, au sens de l'annexe A.2.

En d'autres termes :

$$Prix\ d'acquisition_0 = \sum_{t=1}^{T} VA_{0,TA}(F_t)$$

où :

— $Prix\ d'acquisition_0$ est le prix d'achat de l'obligation au temps 0 ;

— T est la maturité de l'obligation ;

— $VA_{0,TA}(F_t)$ est la valeur actuelle au taux actuariel (TA) et au temps 0 du flux ayant lieu au temps t.

Une fois ce taux déterminé, il est possible de déterminer les amortissements. Pour chaque année i on a :

$$Valeur\ amortie_i = \sum_{t=i}^{T} VA_{i,TA}(F_t)$$

où $VA_{i,TA}(F_t)$ est la valeur actuelle au taux actuariel (TA) et au temps i du flux ayant lieu au temps t. Cette dernière, une fois calculée, permet de déterminer la surcôte/décôte de

(13). L'obligation peut être achetée en cours de vie. Le temps d'achat reste l'origine temporelle.

l'obligation (S/D_i) pour chaque année i. Il s'agit du gain ou perte de valeur de l'obligation d'une année à une autre. Précisément, $\forall i \in \{1, ..., N\}$:

$$S/D_i = Valeur\ amortie_i - Valeur\ amortie_{i-1}$$

Enfin, après avoir calculé le coupon couru (CC_0) au temps 0 de l'obligation (voir A.2.2), on obtient la valeur comptable de l'obligation au temps N :

$$FAV_N = \underbrace{Prix\ d'acquisition_0 - CC_0}_{\text{Prix pied de coupon (14)}} + \sum_{i=1}^{N} S/D_i$$

où :

— FAV_N est la valeur comptable au temps N de l'obligation après prise en compte de l'amortissement ;

— CC_0 est le coupon couru au temps 0 de l'obligation calculé selon la méthode de l'annexe A.2.2.

L'amortissement linéaire

Pour l'amortissement linéaire, la différence entre le prix d'acquisition et le prix de remboursement (le nominal ici) est uniformément amortie jusqu'à maturité.

BNP Paribas Cardif utilise l'amortissement actuariel dans son modèle. L'étude de cette méthode dans notre étude permet de cerner le poids du choc de taux sur le BEL.

2.2.1.2 Le flexing

L'étude des variables ci-dessus souligne déjà le rôle central qu'elles occupent dans le calcul du BEL. Cependant, ce rôle est bien plus important compte tenu d'un mécanisme nommé le *flexing*.

Comme expliqué dans la section 1.2.2, le modèle ALM de Cardif prend en compte une interaction Actif-Passif. L'objectif est de *flexer* l'output déterministe du Passif (les tables DETCF) en prenant en compte l'input stochastique (table ESG) et les actions des assurés (rachats par exemple) ainsi que les décisions du management.

Concrètement, *flexer* les flux du passif revient à les multiplier par des ratios de *flexing*

(14). Voir A.2.2

calculés dans ALM. Trois ratios au total sont calculés pour capter différents effets stochastiques. Deux de ces ratios visent à ajuster les flux du passif (les prestations aux assurés principalement) selon les rachats (totaux dans un cas et partiels dans l'autre) pour simuler l'impact de ce phénomène sur le nombre de contrats. Un dernier ratio est calculé afin de les *flexer* selon la PM (voir 2.2.1.3) du modèle ALM prenant en compte les produits financiers.

Les flux du Passif résultant du *flexing* prennent désormais en compte les aléas du marché, réalisations des projections stochastiques du modèle ALM.

2.2.1.3 Les variables non choquées

Les variables dont il est question dans cette section sont la PM, la PDD, la RC, le PRE et le Cash. Ces variables ne reflètent à priori pas directement les aléas du marché et ne sont donc pas considérées comme variables principales de l'étude. Cependant, il s'agit bien de différentes variables qui devraient influencer le BEL.

Ces variables traduisent certes parfois une partie de l'information sur le BEL contenue déjà dans les variables principales citées ci-dessus. Néanmoins, elles octroient à cette étude une vision plus globale et générale des provisions techniques et peuvent apporter un complément d'information non explicite sur le BEL. Approfondir ces liens ne relève pas des objectifs de ce mémoire.

Toutes ces variables (sauf la PRE) servent cependant au calcul du TRA (2.2.1.1), variable importante dans le mécanisme de participation aux bénéfices. La centralité de ce mécanisme dans le calcul du BEL a été détaillée précédemment et dans la perspective d'obtenir un modèle plus complet, il a semblé préférable de conserver ces variables comme auxiliaires dans l'étude (et par la suite dans le modèle).

La PM

La *PM* ou *Provision Mathématique* représente l'engagement de l'assureur envers son assuré. En terme de montant, elle représente la provision occupant la part la plus importante dans le bilan de l'assureur.

Elle s'exprime comme différence entre la valeur actuelle probable de l'engagement de l'assureur et celle de l'assuré. Il s'agit d'une variable reflétant simplement une partie des prestations dans la formule fermée du BEL.

A tout temps t on a :

$$PM_t = (\ VAP_t(engagements\ assureur) - VAP_t(engagements\ assurés))_+$$

où :

— PM_t est la PM au temps t de la projection dans le modèle ;

— $VAP_t(engagements\ assureur)$ est la valeur actuelle probable au temps t des engagements de l'assureur ;

— $VAP_t(engagements\ assurés)$ est la valeur actuelle probable au temps t des engagements des assurés.

La *PM* est utilisée dans le calcul de l'AMC qui influence directement celui du TRA.

Toutes les variables suivantes (sauf la PRE) sont utilisées dans le calcul des *Produits Financiers Disponibles* qui influence directement celui du TRA.

Les réserves relatives à l'Actif

Les trois variables suivantes sont des réserves explicitement liées à l'Actif et qui sont intégrées en inputs au modèle ALM. Elles sont dotées (15) ou reprises (16) selon l'évolution de l'Actif tout au long de la projection.

La PDD

La *PDD* ou *Provision pour Dépréciation Durable* découle directement de la volatilité du marché. Elle est constituée pour chaque actif du portefeuille dès que la valeur de marché de ce dernier est dépréciée pendant une période jugée comme durable c'est-à-dire 6 mois. L'AMF (17) détermine pour chaque période donnée si le marché est volatile ou non. Une action est alors jugée en dépréciation si sa valeur de marché est inférieure à :

— 80% de sa valeur comptable si le marché est jugé peu volatile ;

— 70% de sa valeur comptable si le marché est jugé volatile.

L'assureur puise dans cette réserve lors de la vente de l'actif qui a permis de la doter.

(15). L'assureur les constitue.
(16). L'assureur puise dans la réserve.
(17). Il s'agit de l'Autorité des Marchés Financiers.

La RC

La *RC* ou *Réserve de Capitalisation* est dotée lors de plus-values liées aux obligations à taux fixe. L'utilité de cette réserve est la prévention de chocs à la baisse (conséquence de mouvements de taux) résultant en moins-values au niveau des obligations et contre l'obligation de vendre ces dernières dans ce cas. Dans la pratique, elle fait partie du fonds propres de l'assureur lors du calcul de solvabilité. Elle appartient en totalité à l'assureur à la fin de la projection.

La PRE

La *PRE* ou *Provision pour Risque d'Exigibilité* est dotée lors de plus-values liées à des actifs autres que les obligations (principalement des actions et des placements immobiliers). La *PDD* étant constituée pour une dépréciation durable, la *PRE* s'y ajoute dans le cas où ces dépréciations sont plus soudaines et imprévisibles. Elle revient à la fois aux assurés et à l'assureur à la fin de la projection.

Cette provision n'a pas été intégrée à la base de données car, comme spécifié précédemment, il ne s'agit pas d'une variable qui rentre en compte directement dans les calculs d'intérêt. Elle a donc été omise de la base de données finale.

Le Cash

Le *Cash* traduit la part d'actif très liquide de l'assureur. Intuitivement, elle représente la richesse réelle que possède l'assureur après s'être constitué toutes les garanties pour couvrir ses risques.

2.3 Construction de la base de données

La construction de la base de données s'est faite de manière à ce qu'elle soit constituée des variables citées ici 2.2.1 (hormis la PRE qui a été omise comme expliqué ici 2.2.1.3) et du BEL correspondant aux différentes combinaisons de ces variables. Cette base de données contient des données tirées des années 2018, 2019 et 2020 de BNP Paribas Cardif.

Afin de construire la base de donées, il a été décidé de procéder en appliquant divers chocs aux trois variables suivantes : la PMVL, la PPE et la courbe de taux.

Les autres variables n'ont pas été choquées et seules les valeurs centrales ont été prises en compte dans la base de données. D'une part, ce choix s'est fait afin d'assurer une certaine simplicité au niveau de la modélisation. D'autre part car l'objectif principal est de cerner l'effet du marché sur le BEL, effet qui n'est pas reflété directement par les variables secondaires. L'effet des variables non choquées sur le BEL est donc surtout mesuré à travers les années.

2.3.1 Les runs Prophet

Prophet a été utilisé afin de calculer le BEL. Chaque workspace Prophet permet de modéliser l'environnement ALS nécessaire pour projeter le BEL d'une année. Trois workspaces ont ainsi été utilisés avec chacun plusieurs runs. Tout run correspond à un scénario donné avec comme inputs des valeurs précises pour toutes les variables étudiées. Un run correspond ainsi à une ligne de la base de données.

2.3.2 Les chocs

Pour chaque année, différents chocs ont été implémentés sur les variables principales : des chocs simples, doubles et triples.

Les chocs simples visent à mesurer la sensibilité du BEL à une seule de ces variables indépendamment des autres. Le choc est appliqué à une seule des trois variables principales alors que les deux autres conservent la valeur centrale.

Les chocs doubles (respectivement triples) visent à mesurer la sensibilité du BEL aux variables principales tout en tenant compte des interactions éventuelles entre ces dernières. Les chocs sont appliqués à deux (respectivement trois) des trois variables pour un même run.

La PMVL

Les chocs appliqués à la PMVL sont : $\pm 1\%$, $\pm 2\%$, $\pm 3\%$, $\pm 4\%$, $\pm 5\%$, $\pm 10\%$ *et* $\pm\, 20\%$. La formule de la PMVL pour un actif i est :

$$PMVL_i = MV_i - FAV_i$$

— MV_i est la valeur de marché de l'actif de la ligne i ;

— FAV_i est la valeur comptable de l'actif de la ligne i.

C'est cette formule qui permet de récupérer la valeur de la PMVL pour chaque année. Concrètement, le tableau *ASSETS_EQUITY* du modèle *ALS* rassemble le portefeuille d'actifs avec la valeur de marché, la valeur comptable et le numéro de pool (ou poche stochastique) correspondant. Ce mémoire se cantonne au pool 1 car il regroupe les actifs du fonds général.

La PMVL du fonds EURO est, pour une année N donnée par :

$$PMVL_N = \sum_{i=1}^{K} PMVL_{i,N} = \sum_{i=1}^{K} (MV_{i,N} - FAV_{i,N})$$

où :

— $MV_{i,N}$ et $FAV_{i,N}$ sont respectivement la valeur de marché et la valeur comptable de l'actif de la ligne i du portefeuille d'actifs de l'année N ;

— K est le nombre d'actifs du pool 1 du portefeuille d'actifs de l'année N ;

— $\forall i \in \{1, ..., K\}$, l'actif de la $i^{\text{ème}}$ ligne appartient au pool 1.

La PMVL choquée s'écrit alors :

$$PMVL_{i,choquée} = PMVL_{i,centrale}(1 + choc)$$

où :

— $PMVL_{i,choquée}$ est la PMVL après choc ;

— $PMVL_{i,centrale}$ est la PMVL initiale non choquée calculée à partir de la formule précédente ;

— *choc* est le choc appliqué sous forme décimale.

On obtient ensuite la valeur de la MV_i pour chaque ligne selon la formule $MV_i = PMVL_i + FAV_i$. On retranscrit ces valeurs MV_i qui correspondent à un choc particulier de la $PMVL$ dans le tableau *ASSETS_EQUITY* et on lance le run. On a alors en résultat la valeur du BEL qui correpond à ce choc de PMVL.

La PPE

Pour des raisons expliquées dans la section 2.2.1.1 de ce mémoire, la variable qui a fait réellement l'objet de chocs dans le mécanisme de PB est la PPE_C. C'est cette variable qui dans le modèle représente la PPE.

Les chocs appliqués à celle-ci sont : ±5%, ±10%, ±15%, ±20%, ±25%, ±30%, ±50%, ±80% *et* ± 100%.

La valeur centrale est récupérée chaque année grâce au tableau *PS_INPUTS_PPE* du modèle *ALS*. Celui-ci rassemble les inputs de la stratégie de Participation aux bénéfices. Ce mémoire se limite au pool 1 car il regroupe les actifs du fonds général et à la PPE contractuelle.

La PPE choquée s'écrit alors :

$$PPE_{i,choquée} = PPE_{i,centrale}(1 + choc)$$

où :

— $PPE_{i,choquée}$ est la PPE après choc ;

— $PPE_{i,centrale}$ est la PPE initiale non choquée calculée à partir de la formule précédente ;

— *choc* est le choc appliqué sous forme décimale.

On obtient ensuite la valeur de la $PPE_{i,choquée}$ qu'on retranscrit dans la première ligne du tableau *PS_INPUTS_PPE* qui correspond au pool 1 et on lance le run. On a alors en résultat la valeur du BEL qui correpond à ce choc de PPE.

La courbe de taux

Les chocs appliqués à la courbe de taux sont de ±10%. Ces chocs n'ont pas été effectués dans le cadre de cette étude mais les résultats d'une autre équipe ont été directement utilisés. En effet, choquer la courbe de taux revient à choquer l'input de la table ESG qui a été inaccessible dans cette étude.

2.3.3 Les limites de la méthode de construction

L'utilisation de *Prophet* pour fournir les données nécessaires à la construction de la base n'était pas optimale mais n'était pas facultative puisque c'était le logiciel de modélisation de Cardif. En effet, ce logiciel n'est pas adapté à la production en masse de différents runs comme nécessaire en Machine Learning. Il est surtout utilisé régulièrement par les équipes pour produire des résultats réguliers mais espacés dans le temps qu'ils soient mensuels, semestriels ou annuels.

Le temps de calcul *Prophet* est assez conséquent (en moyenne deux heures par run) et ces derniers font face à plusieurs obstacles avant d'être effectués. En effet, pour une meilleure gestion des temps de calcul, des priorités par équipe existent afin d'effectuer les calculs importants en premier ce qui handicape parfois les calculs jugés moins prioritaires. Par ailleurs, le logiciel n'exécute que 65 runs au maximum environ en même temps toutes équipes confondues. Un autre obstacle se trouve au niveau du stockage des résultats. Les résultats nécessitent une mémoire conséquente sur le réseau, ce qui pose des soucis au niveau des seuils de mémoire alloués par équipe. Ceci ralentit aussi la production de la base de données car il est nécessaire de lancer seulement un certain nombre de runs à la fois, copier les résultats ailleurs afin de libérer de la mémoire et pouvoir en lancer d'autres.

Ces obstacles en conjonction avec la création manuelle des runs un par un ont été très chronophages au moment de la création de la base de données et ont conduit aussi à quelques bugs résultant en certaines données manquantes. Il a été jugé que le nombre de ces données manquantes était assez faible par rapport à la taille de la base de données. Ces données ont donc été omises au final de la base de données utilisée par la suite.

Finalement, BNP Paribas Cardif se penche sur un transfert de son modèle vers Python dans le futur proche afin de répondre aux problématiques citées ci-haut et de s'adapter plus au besoin de traitement de données massives du monde actuel.

2.3.4 Etude préalable

Cette section du mémoire tente, grâce au repérage effectué précédemment et en détails des variables dans le modèle, de prévoir théoriquement l'effet sur le BEL des chocs effectués comme expliqué dans la section précédente (2.3.1).

La PMVL

Choquer la PMVL revient à choquer la valeur de marché car les valeurs comptables sont supposées fixes (hypothèse simplificatrice par rapport au cas réel expliqué ici 2.2.1.1).

La PMVL et la valeur de marché sont positivement corrélées selon l'équation les reliant. Une augmentation de la MV devraient résulter en une augmentation de la PMVL et donc des résultats de l'assureur (et inversement). L'effet de la PMVL sur les bénéfices de ce dernier ne se résume pas à mais est visuellement explicite dans le calcul du TRA (section 2.2.1.1).

Dans le cas d'une plus-value, l'assureur peut donc servir plus de bénéfices aux assurés selon le

mécanisme de PB expliqué et ses engagements augmentent. Réciproquement, un choc négatif au niveau de la PMVL devrait lui permettre de servir moins de bénéfices aux assurés. Ceci aurait alors pour effet la baisse des engagements de l'assureur. Théoriquement, la PMVL et le BEL sont positivement corrélés.

Par ailleurs, la PMVL devrait influencer la PPE. Un bénéfice trop faible dû à une baisse de la PMVL obligerait l'assureur à puiser dans ses provisions et à utiliser les leviers (2.2.1.1) afin de servir le taux cible ou, à défaut, le taux minimal. Dans ce cas, la PPE diminuerait [(18)]. En revanche, une hausse de la PMVL donnerait l'occasion de doter la PPE.

Finalement, il n'existe pas de lien direct a priori entre la PMVL et la courbe de taux puisque la première concerne principalement les actions alors que la seconde les obligations. Un lien indirect est établi par la suite.

La PPE

Choquer la PPE revient à choquer implicitement les bénéfices dont dispose Cardif. En effet, la PPE est dotée si l'assureur se trouve en mesure de servir un taux supérieur au taux cible (2.2.1.1) et reprise dans le cas contraire.

La PPE devrait donc se montrer corrélée positivement au BEL. D'un autre côté, jouir de bénéfices pourrait résulter d'une performance satisfaisante sur le marché des actions ou des obligations de Cardif. Dans le premier cas, ceci devrait se manifester au niveau de la PMVL qui augmenterait. Dans le second cas, ceci reflèterait peut-être une baisse du taux zéro-coupon d'après A.2.

L'étude préalable de la PPE suggère donc une corrélation positive avec la PMVL et une corrélation négative avec la courbe des taux. La variation de la PPE devrait donc susciter la variation qu'induiraient respectivement les mouvements de la PMVL et de la courbe des taux sur le BEL.

La courbe des taux

D'après A.2, un choc positif sur la courbe des taux (respectivement négatif) valoriserait plus (respectivement moins) une obligation du portefeuille de Cardif. La valorisation des actifs du portefeuille est prise en compte dans l'étape 1 (2.2.1.1) de la participation aux bénéfices.

(18). L'effet de cette diminution est détaillé dans la partie suivante.

Modifier la courbe des taux impacterait alors le TRA et donc l'utilisation des leviers (2.2.1.1).

Concrètement, un choc positif de la courbe des taux aurait pour effet une baisse du TRA et donc des bénéfices de l'assureur. Par la suite, l'assureur verserait moins aux assurés et le BEL diminuerait. De plus, une baisse du TRA engendrerait une reprise de la PPE.

Par contre, un choc négatif de la courbe augmenterait les bénéfices de l'assureur qui verserait plus aux assurés ce qui augmenterait son BEL. Ceci causerait une augmetation du TRA et donc une dotation de la PPE.

En conséquence, l'évolution de la courbe des taux est inversement liée à celle du BEL. Par ailleurs, un mouvement de la courbe des taux altère le TRA ce qui impacte la PPE comme expliqué précédemment (2.3.4). La PPE impactant la PMVL (2.3.4), la courbe des taux a un lien indirect avec la PMVL. La relation réciproque peut aussi être établie.

Le chapitre suivant permettra la modélisation effective du BEL à travers ces chocs.

Les effets d'une baisse de bénéfices

Les trois sections précédentes détaillent à la fois l'effet de la baisse et de l'augmentation des bénéfices. Cependant, il est nécessaire de noter qu'en cas de baisse de bénéfices, indépendamment de la variable la causant, l'assureur puise dans les diverses provisions dont il dispose afin d'absorber ce choc. L'objectif principal reste de pouvoir honorer ses engagements et de rester compétitif. Il est indispensable de prendre en compte l'effet toujours amorti de ces chocs à la baisse qui cause une réelle asymétrie au niveau du BEL.

Afin d'illustrer cette idée et pour plus de clarté : une diminution de 5% des produits financiers de l'assureur résulte certes en une réduction de la part du BEL. Cependant, une augmentation de 5% de ces derniers résulte en une hausse du BEL plus considérable que la réduction précédente.

Cette étude préalable est illustrée et confirmée par la suite sur les données réelles.

Chapitre 3

La modélisation : proxy et backtesting

Ce chapitre rassemble la modélisation effective du BEL à travers les variables explicatives. Il explique les modèles implémentés, présente leurs résultats et leur apporte des explications.

3.1 La modélisation de la courbe des taux : le modèle Nelson Siegel

3.1.1 L'utilié du modèle

Lors de la construction de la base de données, une problématique s'est manifestée : *comment réduire la courbe de taux qu'elle soit centrale ou choquée à quelques variables ?* En effet, utiliser quelques points choisis de la courbe soulèverait plusieurs questions :

— sur quels critères choisirait-on l'omission de certaines points ? Quelle légitimité pour ces critères ?

— combien de points devrait-on garder au final et pourquoi ?

— garder plusieurs points n'allourdirait-il pas la base de données ?

— comment serait-il possible analytiquement d'interpréter le BEL selon certains points seulement de la courbe ?

— l'omission de certains points de la courbe ne causerait-elle pas une vraie perte d'informations ?

3.1.2 La théorie derrière le modèle

Nelson-Siegel est un modèle mathématique qui permet de répondre à cette problématique. Il approxime la courbe de taux zéro-coupon en résumant toute la courbe en quatre paramètres : $\beta_0, \beta_1, \beta_2$ et λ. Ce modèle établit la relation suivante :

$$ZC(\tau) = \beta_0 + \beta_1 \left(\frac{1 - e^{-\lambda\tau}}{\lambda\tau} \right) + \beta_2 \left(\frac{1 - e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau} \right)$$

où :

— $ZC(\tau)$ est le taux zéro-coupon de maturité τ ;

— β_0 est un paramètre dit *long-terme* du modèle. Il traduit le niveau de la courbe ;

— β_1 est un paramètre dit *court-terme* du modèle. Il traduit la pente de la courbe ;

— β_2 est un paramètre dit *moyen-terme* du modèle. Il traduit la courbure de la courbe ;

— λ est le paramètre dit de *decay rate* du modèle. Il traduit la vitesse de décroissance de la courbe.

3.1.3 L'implémentation du modèle et sa pertinence

Le modèle de Nelson-Siegel a été codé sur R du fait de l'existence d'une librairie facilitant l'implémentation de ce dernier. La librairie *YieldCurve* de R permet, par l'intermédiaire d'une fonction *Nelson-Siegel* de cette dernière, de calculer les quatre paramètres cités ci-dessus à l'aide de la courbe de taux seulement en input. Les paramètres d'output calculent alors, pour chaque maturité, un point de la courbe des taux à partir de l'équation précédente et l'approximent alors de cette manière.

L'implémentation du modèle Nelson-Siegel a été effectuée neuf fois au total : trois fois afin de modéliser trois courbes (une centrale et deux choquées à -10% et 10% respectivement) pour chaque année.

Ces neuf modèles permettront d'intégrer les chocs de la courbe des taux à la base de données d'étude dans la suite de ce mémoire. Chaque courbe choquée est représentée par les quatre paramètres Nelson-Siegel du modèle qui lui est relatif.

Avant de procéder à l'utilisation de ces modèles, il a semblé judicieux de visualiser l'efficacité de ce dernier dans un exemple des neuf cas : la courbe centrale correspondant à l'année 2018.

La figure 3.1 superpose pour l'année 2018 les points de la courbe de taux centrale publique

réelle (en rouge) et la courbe générée par le modèle Nelson-Siegel (en bleu) à partir des quatre paramètres. Il est indéniable que la régression semble assez fidèle aux données réelles.

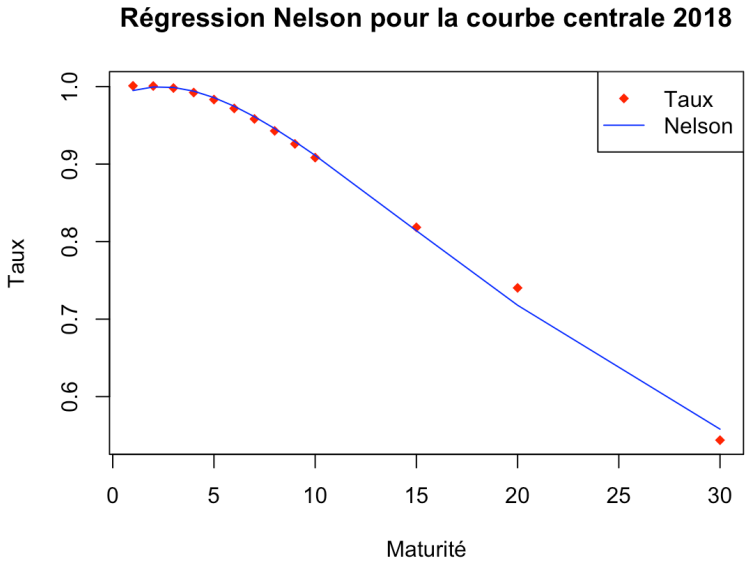

FIGURE 3.1 – L'approximation de Nelson Siegel

La pertinence de ce modèle et sa simplicité en ont fait un choix idéal pour l'étude. Néanmoins, réduire cette courbe à ces quatre paramètres réduit considérablement l'interprétabilité des résultats sur le BEL. Ce modèle reste toutefois idéal pour l'application d'algorithmes de Machine Learning étant donné que ces derniers, en tentant d'établir des liens entre les variables, ne nécessitent pas une prise en compte de leur réelle signification.

A ce stade, la base de données est complète et il est possible de procéder à une exploration de type Machine Learning sur ses données. Une première étape cruciale ici est celle du brassge des données. En effet, il est nécessaire d'éviter un quelconque effet dû à une tendance temporelle dans la base de données qui conduirait à une interprétation erronée du BEL décorrélée des variables explicatives.

3.2 Retour au BEL : l'analyse visuelle

Avant de se lancer dans la modélisation du BEL, il est important de procéder à une analyse visuelle des données. Celle-ci sera réalisée sans prise en compte de l'interaction entre les variables afin de permettre une visualisation plus pure des liens entre le BEL et chacune des variables.

Dans toute cette section, les valeurs visualisées ont été mutilpliées par un coefficient afin de respecter la confidentialité. Les données présentées sont celles de 2018.

3.2.1 Une première analyse relative aux variables explicatives

3.2.1.1 La sensibilité du BEL aux variables principales

L'intérêt se pose tout d'abord, comme énoncé précédemment, sur la relation entre le BEL et les variables principales. Cependant, il est important dans un premier temps d'isoler l'effet de chaque variable sur le BEL afin de l'évaluer correctement. Pour les trois prochaines variables, il a semblé pertinent de procéder ainsi : les lignes de données correspondant à des chocs sur les deux variables qui ne sont pas d'intérêt sont supprimées. Sont conservées les données correspondant à l'éventail de valeurs de la variable d'intérêt et aux valeurs centrales non choquées des deux autres variables.

La PMVL

La première variable d'intérêt est la PMVL. La base de données est modifiée comme expliqué afin de conserver la PPE et les paramètres de la courbe des taux aux valeurs centrales.

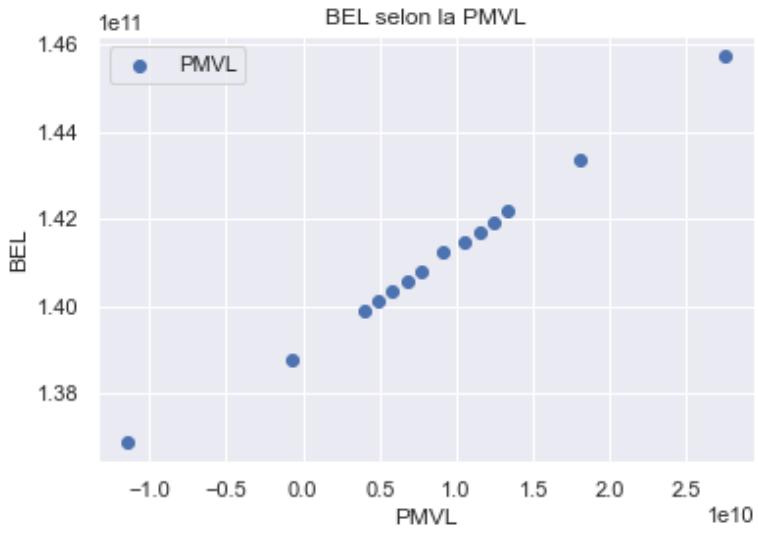

FIGURE 3.2 – Le BEL en fonction de la variation de la PMVL

La figure 3.2 illustre l'analyse préalable stipulée dans la section 2.3.4. En effet, le BEL semble bien présenter une corrélation positive avec la PMVL. La relation semble même quasi-linéaire.

La PPE

La deuxième variable d'intérêt est la PPE. La base de données est modifiée afin de conserver la PMVL et les paramètres de la courbe des taux aux valeurs centrales.

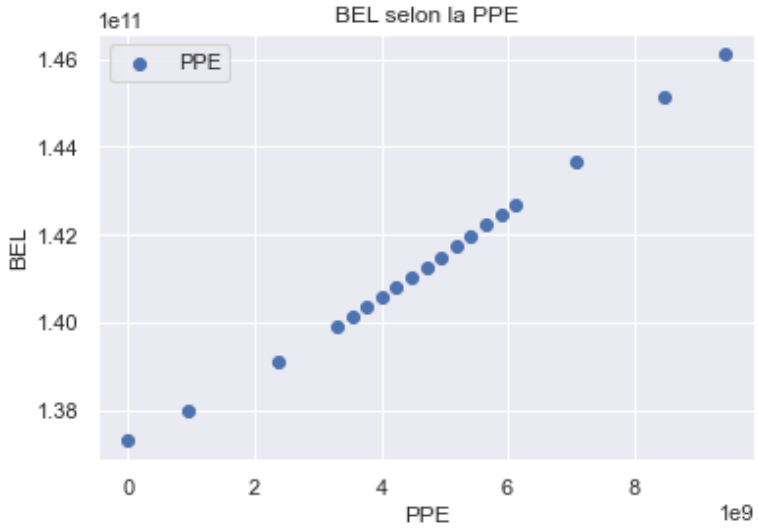

FIGURE 3.3 – Le BEL en fonction de la variation de la PPE

La figure 3.3 illustre l'analyse préalable stipulée dans la section 2.3.4. En effet, le BEL semble bien présenter une corrélation positive avec la PPE. La relation semble même quasi-linéaire.

La courbe de taux

Le questionnement naturel ici au vu des résultats précédents est : existe-t-il une relation visuellement simple de même entre le BEL et certains des coefficients Nelson-Siegel ? Lle même procédé que précédemment a été appliqué, en fixant la PMVL et la PPE à leurs valeurs centrales de 2018.

La figure (figure 3.4) représente le BEL en fonction des quatre paramètres de Nelson Siegel. Les flèches indiquent le choc dont il est question à chaque fois. Il ne semble pas y avoir de relation explicite à première vue entre ces coefficients et le BEL.

Néanmoins, il est possible d'extraire quelques remarques. Tout d'abord le choc est inversement corrélé au BEL. En effet, le choc -10% résulte en le BEL le plus élevé alors que le choc $+10\%$ diminue la valeur du BEL. Ceci est cohérent avec l'analyse préalable (voir 2.3.4).

Une autre remarque est que la courbe de taux centrale ne résulte pas en un BEL parfaitement centré. En effet, un choc positif de la courbe des taux influence plus le BEL qu'un choc négatif de même amplitude. Ceci s'explique également grâce à l'étude préalable effectuée ici (2.3.4).

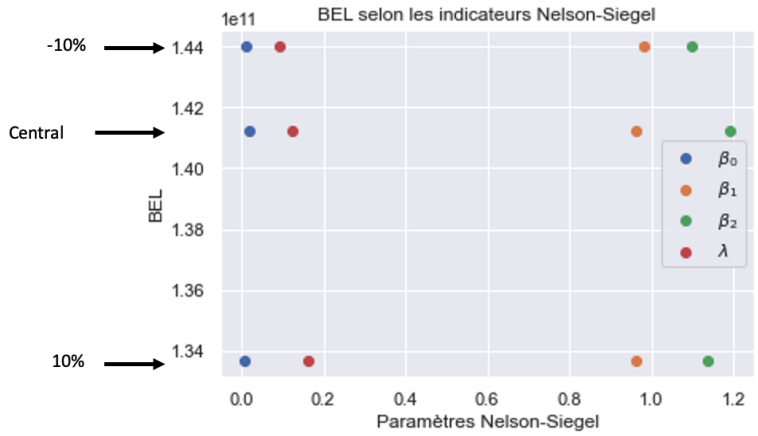

FIGURE 3.4 – Le BEL en fonction des coefficients Nelson-Siegel

Une sensibilité plus ou moins intense

Il peut être intéressant aussi d'illustrer la différence entre les intensités de sensibilité du BEL selon la variable choquée.

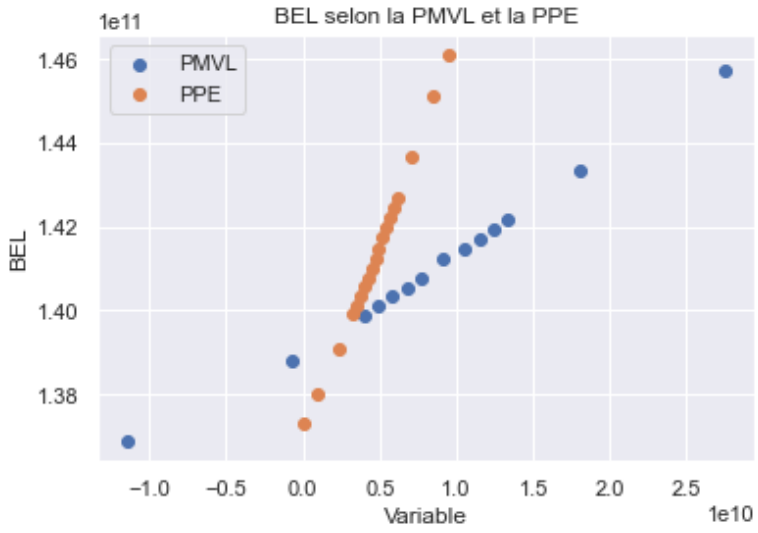

FIGURE 3.5 – Le BEL en fonction de la PMVL et la PPE

Le graphique 3.5 se limite aux chocs unitaires de la PMVL et la PPE seulement afin de souligner la différence de sensibilité du BEL par rapport à ces deux variables. Pour le tracé de la relation BEL-PMVL par exemple et pour les données de l'année 2018, les valeurs de la PPE et des coefficients Nelson-Siegel ont été fixées aux valeurs centrales afin de ne visualiser que l'effet de la PMVL sur le BEL.

Une observation assez claire est que le BEL semble plus sensible à une variation de la PPE que de la PMVL. La pente de la courbe quasi-linéaire qui relie le BEL et la PPE est plus soutenue que celle de la courbe reliant la PMVL et le BEL.

Une tentative d'explication pourrait être établie à partir de l'équation de calcul du TRA (2.2.1.1). Choquer la PPE connoterait un (et résulterait du) choc des bénéfices dans leur globalité et donc de tout le TRA. Ce dernier représente les bénéfices des actions (liées à la PMVL) et des obligations (liées à la courbe des taux). Un mouvement de la PPE devrait donc être plus extrême sur le BEL que celui de la PMVL ou de la courbe des taux.

Afin de préserver la lisibilité des graphiques, il a été décidé de ne pas en représenter un similaire à celui de la figure 3.5 pour souligner cette différence entre la courbe des taux et les autres variables.

Obtenir un graphique similaire à celui de la figure 3.5 afin de souligner la différence de sensibilité du BEL par rapport à la courbe des taux et les autres variables n'est pas possible. En effet, la courbe étant représentée par ses paramètres Nelson Siegel, le graphique ne serait pas aussi démonstratif.

3.2.1.2 L'interaction entre les variables

Les matrices de corrélation peuvent soulever les interactions entre les variables quantitativement. Ces dernières croisent toutes les variables entre elles selon un coefficient de corrélation choisi.

Le calcul de cette matrice s'est fait pour deux coefficients différents qui traduisent deux types de relations différentes : le coefficient de Pearson et celui de Spearman.

Le coefficient de Pearson traduit la relation linéaire qui existe entre deux variables. Un coefficient de Pearson proche de 1 en valeur absolue suggère un fort lien linéaire entre les variables alors qu'une valeur se rapprochant de zéro suggère l'existence d'une faible relation linéaire voire l'absence d'une telle relation.

Pour deux variables X et Y de réalisations $(X_i)_{1\leq i\leq n}$ et $(Y_i)_{1\leq i\leq n}$, le coefficient de Pearson $\rho_P(X,Y)$ s'exprime comme suit :

$$\rho_P(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Le coefficient de Spearman traduit une relation ordinale entre deux variables, une monotonie

entre ces dernières même si elle ne s'exprime pas à la même vitesse pour les deux. Il est égal au coefficient de Pearson appliqué aux rangs $rg(X_i)$ et $rg(Y_i)$ des échantillons X_i et Y_i.

Ce mémoire ne s'attardera pas sur l'étude de la matrice de corrélation de Spearman puisqu'elle a conduit aux mêmes conclusions que celles suggérées par la matrice de Pearson. La matrice de corrélation de Pearson est illustrée par la figure 3.6.

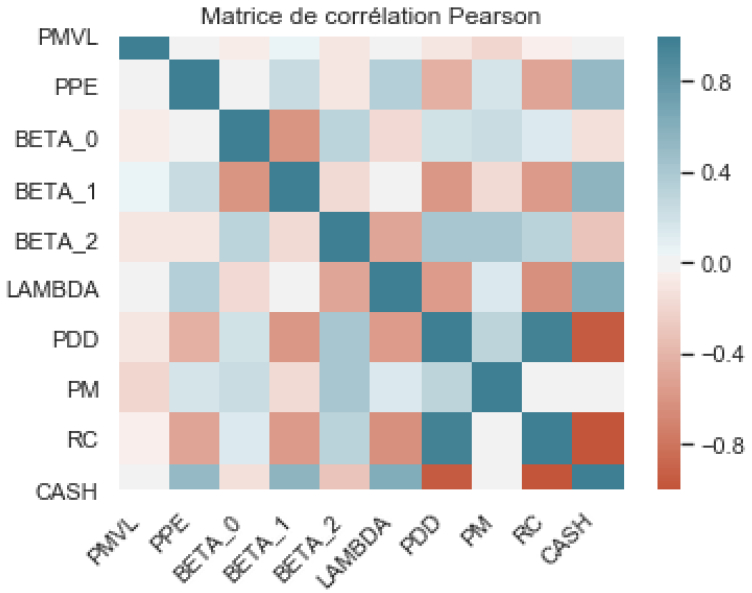

FIGURE 3.6 – La matrice de corrélation de Pearson

On remarque que la RC est très corrélée avec la PDD et le Cash et naturellement que le Cash est très corrélé avec la PDD et la RC. La présence des variables RC et Cash peut donc biaiser les modèles. Il a été décidé de les ôter de l'ensemble des variables explicatives.

Les prochaines sections tentent de modéliser le BEL à travers les variables explicatives conservées : la PMVL, la PPE, les coefficients de Nelson-Siegel, la PM et la PDD.

3.2.2 Une analyse du BEL de la base de données

Le boxplot du BEL

Les performances des algorithmes de Machine Learning sont généralement médiocres lorsque les données sont *skewed*. L'étude de la distribution du BEL s'impose donc avant tout.

Tout d'abord, le boxplot (figure 3.7) [(1)] présente une régularité satisfaisante du comporte-

(1). Les valeurs ont été à nouveau multipliées par un certain coefficient par respect de la confidentialité.

ment de la variable à expliquer. Il n'est a priori pas nécessaire de procéder à une étude de type *Théorie des valeurs extrêmes* afin de lisser les valeurs aberrantes ou trop extrêmes dees données.

FIGURE 3.7 – Le boxplot du BEL

La distribution du BEL

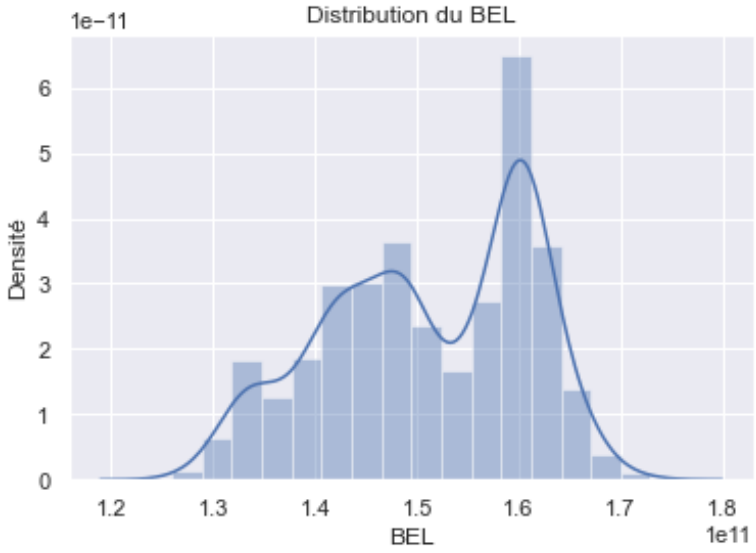

FIGURE 3.8 – La densité du BEL

Une figure complémentaire semble intéressante afin de mettre en lumière un autre effet. La

distribution du BEL (voir figure 3.8) [2] met en valeur une concentration des valeurs du BEL favorable aux grandes valeurs. Cet effet est visible pour les grandes valeurs du BEL et ne l'est pas, par contre, pour les valeurs faibles du BEL. Ceci peut être interprété comme suit :

— D'un côté, le processus de participation aux bénéfices privilégie une distribution profitable aux assurés pour rester concurrentiel et éviter les rachats dynamiques. Si une variation des variables principales favorise l'augmentation du BEL alors l'assureur n'a pas intérêt à contrer cette effet ;

— D'un autre côté, si les variations des variables principales conduisent à une baisse des bénéfices et, par la suite, du montant à servir aux assurés alors l'assureur utilise les leviers (2.2.1.1) à sa disposition afin d'absorber ce risque ;

— Enfin, un BEL élevé est, globalement, synonyme d'une performance positive sur le marché des actifs du BEL et l'assureur peut se permettre de verser plus aux assurés. Ceci rejoint les deux idées précédentes. Cependant, une performance faible ou négative sur le marché oblige l'assureur à utiliser aussi d'autres provisions que les leviers du point précédent. Il s'agit notamment de provisions telles que la RC et la PDD (2.2.1.3) qui permettent elles aussi d'amortir les chocs à la baisse et de maintenir un niveau satisfaisant du BEL.

3.3 Automatisation du proxy

Cette partie détaille la recherche d'un modèle de type *Machine Learning* afin de modéliser la relation entre le BEL et les variables explicatives. L'entrainement a reposé sur de l'apprentissage supervisé (puisque les valeurs du BEL sont disponibles) afin de tester plusieurs modèles sur la base de données. Tous les modèles sont testés sur les données des années 2018 et 2019. Les données de l'année 2020 seront utilisées ultérieurement dans un objectif de Backtesting du modèle choisi dans cette section.

3.3.1 Les modèles

Afin d'identifier le modèle optimal, il a fallu tester une multitude de modèles. Au vu des relations parfois quasi-linéaires entre le BEL et certaines des variables explicatives principales, les modèles tels que la *Régression Linéaire* ou les *GLM* (Gamma et Gaussian) [3] semblent pertinents à tester en premier lieu.

(2). Voir (1)
(3). Generalized Linear Models

Par la suite, des modèles moins simples ont été implémentés tels que des *Decision Trees*, *Adaboost*, *Gradient Boosting* et *Random Forest*. Ceux-ci ont été jugés utiles à coder pour diverses raisons :

— La relation entre les paramètres de Nelson-Siegel et le BEL semble plus complexe et des modèles plus sophistiqués paraissent plus adéquats pour l'identifier ;

— Une relation non-linéaire et plus délicate peut se cacher derrière les liens entre les données ;

— Une volonté d'atteindre une plus grande précision motive l'implémentation de modèles plus élaborés.

Pour chaque modèle plusieurs indicateurs ont été calculés afin de permettre de les comparer plus tard et juger de leur pertinence : des scores du modèle, la racine de l'erreur quadratique moyenne, l'erreur en pourcentage par rapport au BEL moyen, le BIC (pour certains modèles) et le vecteur d'erreur en pourcentage par rapport à tous les BEL. Le cadre théorique définissant ces indicateurs a été présenté ici 3.3.2.

3.3.1.1 Le Prétraitement des données

Avant de procéder en appliquant divers algorithmes de Machine Learning, il est primordial de scinder la base de données [(4)] en deux parties : une base dite *train* sur laquelle l'algorithme s'entrainera et une autre dite *test* qui permettra de valider ou non temporairement le modèle. Cette séparation est aléatoire. Les algorithmes suivants ont donc été entrainés sur la base *train* qui constitue 70% des données.

Dans la suite, après avoir entrainé l'algorithme sur le jeu de données *train*, plusieurs indicateurs et graphiques ont été générés à partir du jeu de données *test*. Les valeurs des axes ont parfois été omises par respect de la confidentialité.

Pour la suite dans ce mémoire, la variable à expliquer est notée Y et l'ensemble des variables explicatives est symbolisé par X .

3.3.1.2 La Régression Linéaire

La Régression Linéaire est le premier et le plus simple modèle ajusté aux données. Il cherche à établir une relation de la forme :

(4). La base utilisée ici est composée des données 2018 et 2019 seulement.

$$Y = X\beta + \varepsilon$$

où :

— Y est de taille nx1 ;

— $X \in \mathbb{M}_{(n,p)}(\mathbb{R})$ et de rang plein ;

— β est de taille px1 et ε est de taille nx1 ;

— $\mathbb{E}(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$ qui est inconnue et les ε_i sont iid.

Intuitivement, Y *doit rester en moyenne autour de* $X\beta$. L'algorithme estime alors β qui permet d'avoir au mieux cette relation. La solution, $\hat{\beta}$, correspond au β minimisant la quantité $\|Y - X\beta\|^2$.

Visualiser cette régression n'est pas raisonnable puiqu'elle repose sur une multitude de variables explicatives. Une manière simple, cependant, de visualiser la performance de cette dernière est de représenter les Y_i en fonction des $\hat{Y}_i := \left(X\hat{\beta}\right)_i$ pour la base de données *test*. Si l'adéquation entre ces valeurs est satisfaisante alors les points doivent suivre le tracé de la fonction $f : x \mapsto x$. La figure 3.9 représente cette adéquation [5].

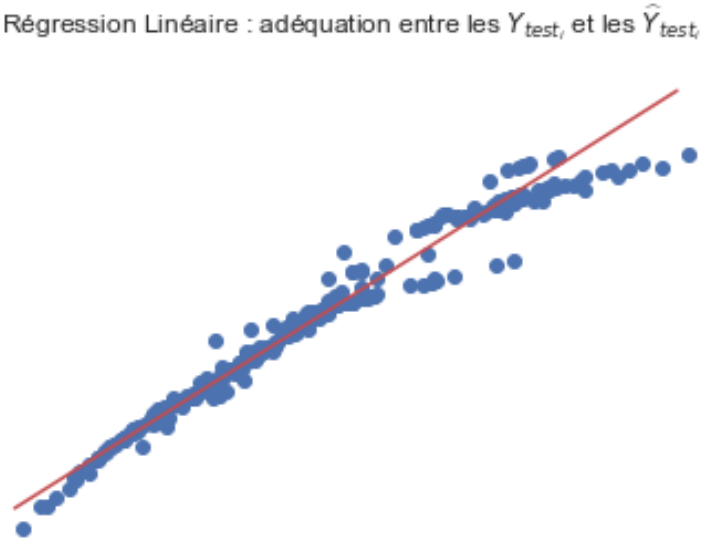

FIGURE 3.9 – Régression Linéaire : l'adéquation entre les Y_i et les $\hat{Y}_i$

La figure 3.9 montre que le modèle est assez correct. Les valeurs prédites de la base *test* correspondent quasiment aux valeurs réelles. Cependant, le modèle semble assez simple et rigide

(5). Ce graphique s'appuie sur le jeu de données *test*.

et des écarts sont observés. La complexité de la relation entre les variables et le BEL ne semble pas totalement captée par le modèle.

Le tableau suivant récapitule les résultats principaux permettant d'évaluer la performance de ce modèle et de le comparer par la suite à d'autres modèles. Le calcul de ces indicateurs ainsi que leur utilité seront énoncés dans la partie 3.3.2. La comparaison entre les valeurs des indicateurs des différents modèles sera effectuée par la suite (section 3.3.2.1).

Performance de la Régression Linéaire	
Indicateur de performance	Valeur
RMSE	1263372567
Score	0.971
BIC	11747.218
Moyenne du score de cross-valisation	0.966

FIGURE 3.10 – Les principaux indicateurs de performance de la Régression Linéaire

Il est clair que la Régression Linéaire capte en partie les liens entre les variables explicatives et le BEL. Les modèles suivants, plus complexes, ont été étudiés pour les raisons citées ici 3.3.1.

3.3.1.3 Les Modèles Linéaires Généralisés ou GLM

Les *Modèles Linéaires Généralisés* sont, comme leur nom l'indique, une généralisation du modèle linéaire.

Il s'agit de supposer le cadre statistique suivant :

$$\begin{cases} Y \text{ suit une loi parmi la famille exponentielle} \\ \exists\, g : \mathbb{R}^{\mathbb{N}} \mapsto \mathbb{R}^{\mathbb{N}} \text{ monotone}, \beta \in \mathbb{R}^{\mathbb{N}}, \text{ tels que } g(\mathbb{E}[Y|X]) = X\beta \end{cases}$$

La famille de loi exponentielle est un ensemble de lois de probabilité. X variable aléatoire continue de densité $f_{X,\theta}$ suit une loi de la famille exponentielle si :

$$f_{X,\theta}(x) = a(\theta)b(x)e^{\eta(\theta).T(x)}$$

où $a(\theta),\ b(x),\ \eta(\theta)$ et $T(x)$ sont bien définis.

Calibrer un *GLM* aux données revient à supposer que :

— Y suit une certaine loi de la famille exponentielle. Cette hypothèse est à vérifier grâce à

un test statistique ;

— Y est relié au prédicteur linéaire $X\beta$ à travers la fonction monotone g appelée *fonction de lien*.

L'objectif des deux sections suivantes est de calibrer deux *GLM* sur les données. Il s'agit de tester l'adéquation du BEL aux deux lois suivantes : la loi Gamma et la loi Normale. Ceci est effectué, théoriquement, à l'aide du test de Kolmogorov B. Pratiquement, le package *scikit-learn* de *Python* effectue ce test automatiquement.

Gaussian GLM

Dans un *Gaussian GLM*, il est supposé que le cadre précédent est vérifié et que la loi de la famille exponentielle dont il est question est la loi Normale.

De même que pour la régression linéaire, visualiser cette régression n'est pas raisonnable puiqu'elle repose sur une multitude de variables explicatives. Le processus de visualisation choisi est alors le même. La figure 3.11 représente l'adéquation entre les Y_i et les $\hat{Y}_i$ [6].

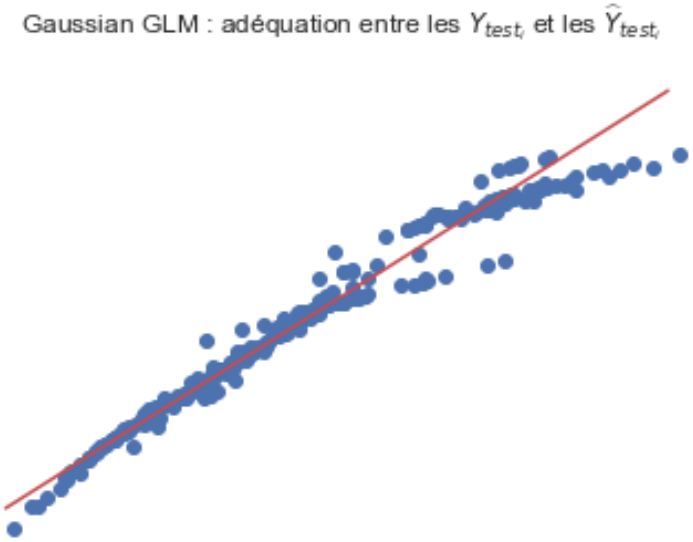

FIGURE 3.11 – Gaussian GLM : l'adéquation entre les Y_i et les $\hat{Y}_i$

La figure 3.11 montre que le modèle est assez correct. Les valeurs prédites de la base *test* correspondent bien aux valeurs réelles. Cependant, le modèle ne semble pas présenter d'amélioration considérable par rapport à la régression linéaire. Par ailleurs, la complexité de la relation entre les variables et le BEL ne semble pas totalement captée par le modèle.

(6). Ce graphique s'appuie également sur le jeu de données *test*.

Le tableau suivant récapitule les résultats principaux permettant d'évaluer la performance de ce modèle et de le comparer par la suite à d'autres modèles. Le calcul de ces indicateurs ainsi que leur utilité seront énoncés dans la partie 3.3.2. La comparaison entre les valeurs des indicateurs des différents modèles sera effectuée par la suite (section 3.3.2.1).

Performance de la Gaussian GLM	
Indicateur de performance	Valeur
RMSE	1263372594
Score	—
BIC	$1.25x10^{21}$

FIGURE 3.12 – Les principaux indicateurs de performance de la Gaussian GLM

Gamma GLM

Dans un *Gamma GLM*, il est supposé que le cadre précédent est vérifié et que la loi de la famille exponentielle dont il est question est la loi Gamma.

De même que pour les deux modèles précédents, visualiser cette régression n'est pas raisonnable. Le processus de visualisation choisi est alors le même. La figure 3.13 représente l'adéquation entre les Y_i et les $\hat{Y}_i$ (7).

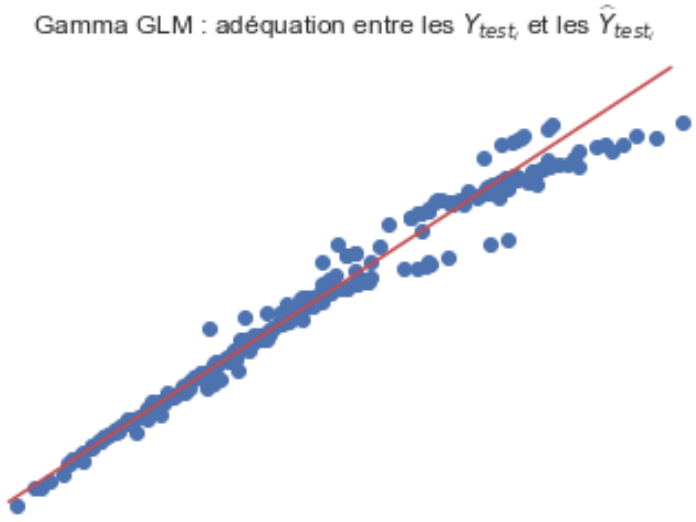

FIGURE 3.13 – Gamma GLM : l'adéquation entre les Y_i et les $\hat{Y}_i$

(7). Ce graphique s'appuie également sur le jeu de données *test*.

La figure 3.13 montre que le modèle est assez correct. Les valeurs prédites de la base *test* correspondent quasiment aux valeurs réelles.

Par ailleurs, le modèle semble présenter une légère amélioration par rapport aux deux modèles précédents. Cette légère amélioration pourrait être expliquée visuellement à partir de la figure représentant la distribution du BEL (voir 3.8). En effet, une certaine asymétrie dans la distribution est visible. Une explication est apportée à cette tendance ici 3.2.2. Une distribution de type Gamma peut présenter une asymétrie et permettre plus de flexibilité par rapport à une distribution Normale. Toutefois, le modèle n'est pas idéal et peut considérablement être amélioré.

Le tableau suivant récapitule les résultats principaux permettant d'évaluer la performance de ce modèle et de le comparer par la suite à d'autres modèles. Le calcul de ces indicateurs ainsi que leur utilité seront énoncés dans la partie 3.3.2. La comparaison entre les valeurs des indicateurs des différents modèles sera effectuée par la suite (section 3.3.2.1).

Performance de la Gamma GLM	
Indicateur de performance	Valeur
RMSE	1168727538
Score	—
BIC	-4187

FIGURE 3.14 – Les principaux indicateurs de performance de la Gamma GLM

3.3.1.4 Le Gradient Boosting

Plusieurs algorithmes de type *Decision Trees* ont été implémentés sur les données. Cependant, et pour ne pas gêner la lisibilité de ce mémoire, il a semblé préférable de ne détailler ici que la théorie derrière ainsi que les résultats de l'algorithme qui s'est avéré être le plus performant d'entre eux : le *Gradient Boosting*.

Une définition du cadre des algorithmes de type *Decision Trees* est cependant nécessaire. Pour plus de détails voir B.2.

L'algorithme de *Gradient Boosting* est un algorithme d'apprentissage supervisé de type *Decision Trees* qui améliore constamment des *weak learners*[8] encore assez simples en allouant un poids plus conséquent aux erreurs à chaque itération.

(8). Cette expression désigne les algorithmes peu performants.

L'objectif de chaque itération est la minimisation de la *RMSE*. A chaque pas k, il faut chercher h_{k-1} de manière à avoir $h_{k-1}(X_i) = Y_i - L_{k-1}(X_i)$ ou de manière équivalente :

$$L_k(X_i) = L_{k-1}(X_i) + h_{k-1}(X_i) = Y_i$$

où :

— L_{k-1} est un *weak learner* que l'algorithme cherche à améliorer ;

— h_{k-1} est un nouvel estimateur permettant cette amélioration ;

— X_i et Y_i sont respectivement les valeurs des variables explicatives et la variable à prédire.

Les $Y_i - L_{k-1}(X_i)$ sont appelés résidus. Par ailleurs, le *Gradient Boosting* est un algorithme dit de *Boosting* comme son nom l'indique. Ceci signifie que chaque *learner* tente de *booster* ou améliorer la prédiction du *learner* précédent. Cette technique est à confronter au *Bagging* qui expliquée dans l'annexe (voir B.2).

Le processus de visualisation choisi pour visualiser les résultats de cet algorithme est le même que précédemment. La figure 3.15 représente l'adéquation entre les Y_i et les $\hat{Y}_i$ [9].

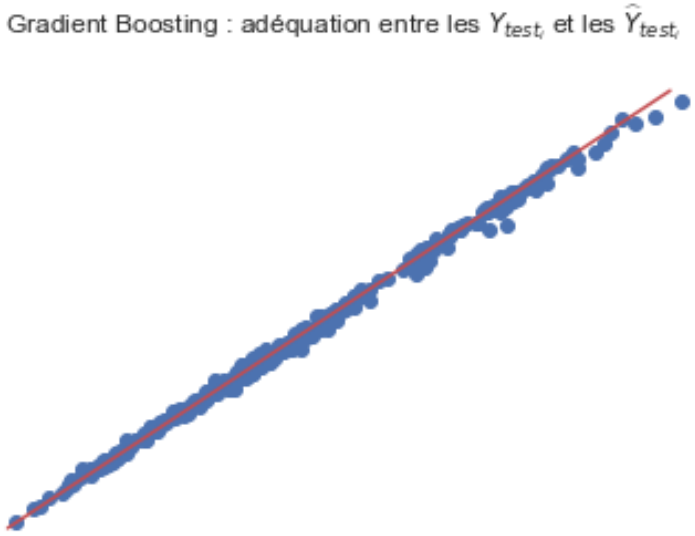

FIGURE 3.15 – Gradient Boosting : l'adéquation entre les Y_i et les $\hat{Y}_i$

La figure 3.15 montre que le modèle est très performant. Les valeurs prédites de la base *test* correspondent quasiment parfaitement aux valeurs réelles.

Par ailleurs, le modèle semble présenter une nette amélioration par rapport aux trois modèles

(9). Ce graphique s'appuie également sur le jeu de données *test*.

précédents. Cette amélioration pourrait être expliquée tout simplement par la complexité de l'algorithme.

Le tableau suivant récapitule les résultats principaux permettant d'évaluer la performance de ce modèle et de le comparer par la suite à d'autres modèles. Le calcul de ces indicateurs ainsi que leur utilité seront énoncés dans la partie 3.3.2. La comparaison entre les valeurs des indicateurs des différents modèles sera effectuée par la suite (section 3.3.2.1).

Performance du Gradient Boosting	
Indicateur de performance	Valeur
RMSE	405440392
Score	0.999
BIC	—
Moyenne du score de cross-valisation	0.997

FIGURE 3.16 – Les principaux indicateurs de performance du Gradient Boosting

Ce modèle représente bien tous les autres modèles de type *Decision Trees* implémentés et dont la théorie est détaillée ici B.2 et qui ont tous été particulièrement performants pour prédire les données.

La section suivante présentera les différents indicateurs de performance retenus et comparera les différents modèles de manière plus détaillée afin de déterminer un modèle optimal.

3.3.2 La performance des modèles

3.3.2.1 Les indicateurs de performance

Les indicateurs de performance permettent de quantifier la qualité d'un modèle. Voici les principales méthodes utilisées par la suite dans ce mémoire.

Le score

Cet indicateur est calculé automatiquement dans Python. Un score proche de 1 reflète une excellente performance du modèle alors qu'un score proche de 0 signifie que le modèle ne réussit pas à prédire les données.

Le BIC

Le *BIC* ou *Bayesian Information Criterion* est un critère permettant de juger de la qualité d'un modèle. Il s'agit d'un indicateur s'inspirant fortement de l'AIC[10]. Ces deux critères permettent de juger de la parcimonie d'un modèle et le pénalisent dans le cas où il repose sur un grand nombre de paramètres. L'AIC s'exprime comme suit :

$$AIC = 2k - 2ln(L)$$

où k est le nombre de paramètres du modèle et L le maximum de sa fonction de vraisemblance.

Le critère *BIC* prend également en considération la taille de l'échantillon. En effet, il est raisonnable qu'un échantillon plus grand nécessite un modèle plus complexe.

Le *BIC* s'exprime comme suit :

$$BIC = kln(n) - 2ln(L)$$

où n est le nombre d'observations, k le nombre de paramètres du modèle et L le maximum de sa fonction de vraisemblance.

La RMSE

La *RMSE* ou *Root Mean Squared Error* évalue la tendance à l'erreur d'un modèle. Concrètement, l'indicateur mesure l'écart entre les valeurs prédites par un modèle et les valeurs réelles. La formule est la suivante :

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}$$

Le boxplot des RMSE relatives

(10). Akaike Information Criterion.

Afin de pouvoir représenter le boxplot des RMSE relatives, il est nécessaire de calculer la RMSE (3.3.2.1). En notant $Y_{boxplot}$ le vecteur contenant les valeurs que représente ce boxplot, la formule suivante est celle permettant de le calculer :

$$Y_{boxplot_i} = \frac{RMSE}{Y_i}$$

où :

— $Y_{boxplot_i}$ est la i-ème composante du vecteur $Y_{boxplot}$;

— $RMSE$ est l'indicateur calculé ici 3.3.2.1 ;

— Y_i est la i-ème composante du vecteur contenant les valeurs réelles du BEL.

Ce boxplot représente donc, intuitivement, l'écart, en pourcentage, que constitue l'erreur du modèle par rapport à chaque valeur réelle à prédire. Il est donc possible de visualiser l'étendue de l'erreur ainsi que sa médiane.

La cross-validation

La *cross-validation* ou *validation croisée* est une technique d'apprentissage évaluant la fiabilité d'un modèle. La technique de *cross-validation* utilisée dans ce mémoire divise la base de données en $k = 5$ blocs de tailles égales. Pour $k \in \{1, ..., 5\}$, il s'agit de calculer le score en utilisant le k-ème bloc comme base de données *test* et les quatre autres comme *train*.

Cette méthode garantit que la base de données *train* choisie n'est pas particulière par rapport au reste et que le modèle n'effectue pas de surapprentissage (3.4).

3.3.2.2 Le modèle choisi

Tous les modèles ont réussi à modéliser, du moins en partie, le BEL de façon correcte et avec, relativement, assez peu d'erreurs. Il est cohérent que plusieurs de ces modèles puissent modéliser si bien le BEL pour diverses raisons :

— les valeurs du BEL dans la base de données sont assez uniformes dans le sens où il n'y a pas de valeurs aberrantes et que les valeurs ne présentent par de gros écarts les unes par rapport aux autres (voir 3.7).

— la relation à modéliser entre le BEL et deux des variables principales (la PMVL et la PPE) ne semble pas s'écarter énormément d'une relation linéaire. Elle devrait donc être

assez simple et homogène pour permettre une bonne modélisation.

L'objectif de cette section est de rassembler toutes les informations de la section précédente et de choisir le modèle optimal.

A première vue et graphiquement, le modèle *Gradient Boosting* a l'air de restituer la prédiction la plus fidèle des données (voir la figure 3.15). Une autre visualisation des erreurs permettant de comparer ces modèles plus clairement est présentée ci-dessous afin de s'assurer de la validité de ce modèle par rapport aux autres.

Les boxplots suivants (figure 3.17) représentent, pour chaque modèle, la racine de l'erreur quadratique moyenne (la RMSE) par rapport à toutes les valeurs réelles du BEL à prédire. Ils permettent donc de voir l'étendue de l'écart de l'erreur par rapport aux valeurs réelles en pourcentage [(11)]. Ce graphique confirme la pertinence de la quasi-totalité des modèles puisque pour tous ces derniers l'erreur est inférieure au pire des cas à un écart de 2%.

En se référant à nouveau au graphique 3.17, il est facile de noter que le modèle *Gradient Boosting* semble être le modèle optimal pour répondre à notre problématique (suivi de peu par le *Random Forest* puis les *Decision Trees*).

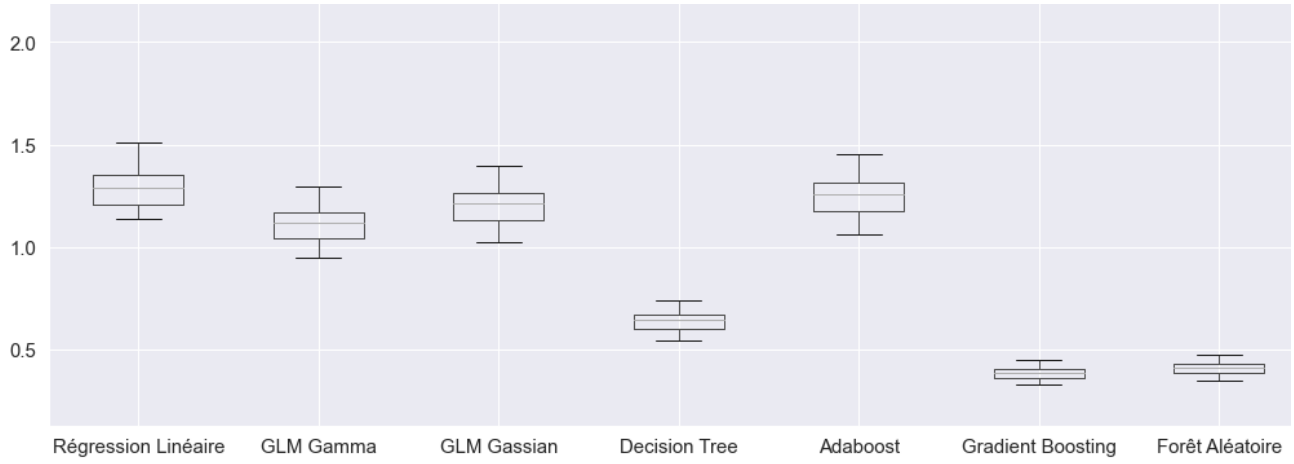

FIGURE 3.17 – Les boxplots de l'erreur quadratique moyenne en pourcentage par rapport aux valeurs réelles des différents modèles

Ce boxplot confirme également les observations des sections précédentes à savoir que les modèles linéaire et linéaires généralisés présentent une performance correcte pour leur simplicité et leur interprétabilité.

Si le choix d'un modèle simple est le critère principal alors le modèle optimal est le *GLM Gamma*. En effet, il présente certes un *RMSE* plus élevé que le *Gradient Boosting* mais il est

(11). Voir 3.3.2.1 pour plus de détails à propos du calcul de ces boxplots.

nettement plus simple et plus interprétable. Par rapport aux autres modèles linéaires, il présente moins d'erreurs et même si cette différence n'est pas considérable, son *BIC* est nettement inférieur. Avec moins de paramètres, il capte mieux les relations entre les données.

Cependant, l'objectif de ce mémoire est, avant tout, de présenter le meilleur modèle de prédiction et l'aspect quantitatif prime sur l'interprétabilité. L'algorithme choisi est alors celui du *Gradient Boosting*. Toutefois, il est intéressant de noter que les modèles linéaires, de part leur interprétabilité, restent un choix raisonnable quantitativement dans le cas où une analyse qualitative est aussi due.

Dans la suite, le modèle utilisé et conservé pour cette étude est celui du *Gradient Boosting*.

3.4 La validation du modèle

Le modèle choisi semble assez performant. Néanmoins, il est nécessaire de procéder à une ultime étape permettant de le valider pleinement : la validation du modèle. Cette étape évalue sa tendance au surapprentissage, le teste sur un nouveau jeu de données et confronte les attentes aux résultats réels.

Le surapprentissage ou *overfitting*

Cette section questionne la tendance du modèle au surapprentissage. Un modèle qui fait du surapprentissage ou de l'*overfitting* calque le comportement des données au lieu d'en extraire un motif global. C'est comme si l'algorithme « se rappellait » des données au lieu de les « apprendre ». Dans ce cas, il apprend même le bruit et la particularité des données et présente une performance mauvaise face à un nouveau jeu de données.

Vue la performance des modèles dans la section précédente, il est important de se questionner à propos d'un éventuel surapprentissage. Il est possible de stipuler, a priori, que ces modèles ne comportent normalement pas de surapprentissage pour les raisons suivantes :

— La base de données est assez homogène. En effet, la variable à prédire ne présente pas de valeurs aberrantes ou trop extrêmes. Que les algorithmes prédisent bien les données n'est pas forcément surprenant ;

— Des modèles assez simples tels que la *Régression Linéaire* ou les *GLM* présentent de très bonnes approximations aussi. Or, le surapprentissage découle surtout de la complexité du modèle ;

— Les modèles présentent des scores excellents à la fois pour la base d'apprentissage et de validation. Dans un cas de surapprentissage, le modèle apprend surtout les données d'apprentissage et prédit mal les données de validation ;

— Tous ces modèles présentent des moyennes de scores convenables de cross-validation.

Pratiquement, tous les modèles précédents présentent un score moyen de cross-validation très correct. Il est donc possible de procéder sans inquiétude.

Le Backtesting

L'objectif du Backtesting est de tester un modèle calibré exclusivement sur un jeu de données A sur un nouveau jeu de données B afin d'en évaluer la pertinence et de détecter pratiquement un quelconque *overfitting*.

Le modèle choisi dans cette étude a été calibré sur des données 2018 et 2019. Le backtesting sera effectué sur les données de l'année 2020. Comme énoncé le modèle *Gradient Boosting* est conservé et utilisé ici.

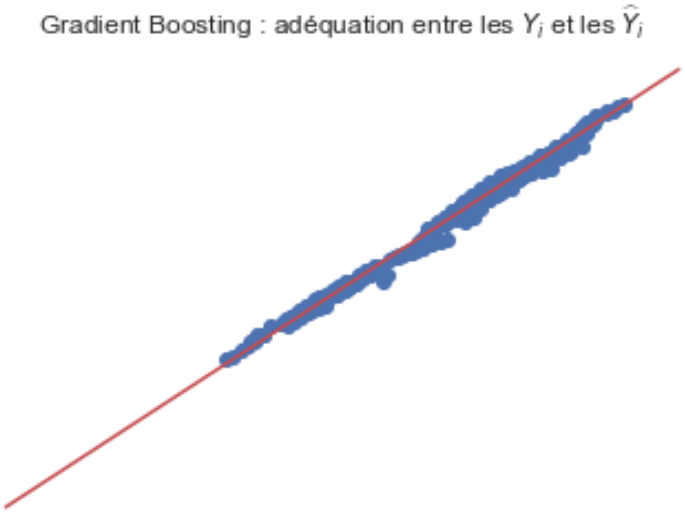

FIGURE 3.18 – L'adéquation entre les valeurs réelles et les valeurs prédites pour le Gradient Boosting pour les données 2020

Le graphique précédent 3.18 représente en abscisses les Y_i ou valeurs réelles du BEL de la base de validation et, en ordonnées, les $\hat{Y}_i$ ou valeurs prédites grâce au modèle de *Gradient Boosting*. Comme pour les graphiques précédents, plus les points s'alignent donc sur une droite parfaite, plus les prédictions sont excellentes. Il est facile de voir, dans le cas du modèle choisi

de *Gradient Boosting*, que le modèle génère de très bonnes prédictions.

Ceci confirme donc l'adéquation du modèle trouvé. Il ne présente pas de surapprentissage puisqu'il est tout aussi performant sur de nouvelles données.

3.4.1 Résultats et discussions

Dans cette étude, l'algorithme retenu de type *Gradient Boosting* est très performant. Cet outil devrait permettre un excellent proxy du BEL d'autant plus qu'il a été calibré sur un grand nombre de données.

La performance des modèles a, autant que possible été confrontée aux attentes théoriques selon les variables utilisées.

Toutefois, plusieus choix ont été effectués afin de mener cette étude qui pourraient être remis en question. En effet, certaines variables, après une analyse plus approfondie du modèle de Cardif, pourraient être plus performantes pour prédire le BEL que les variables choisies. De plus, l'utilisation d'une approche de type *Machine Learning* est un choix comme un autre.

D'autres méthodes ainsi que d'autres variables pourraient apporter une dimension intéressante à l'étude de la sensibilité du BEL. Lextension de cette étude à d'autres indicateurs tels que le SCR serait intéressante.

Conclusion

Cette étude a permis de répondre à la problématique de besoin d'efficience de calcul du BEL, indicateur certes indispensable au reporting mais très précieux pour l'assureur-même. En effet, il peut aider à la prise de décisions d'investissement, permet de se projeter et d'évaluer la santé de son bilan. C'est un outil incontournable de la gestion des risques pour l'assureur.

Plus particulièrement, la mission principale d'évaluer les effets sur le BEL des aléas du marché a orienté l'étude dans une voie presque instinctive : celle d'identifier les variables reflétant les variations du marché, les choquer pour mimer ces aléas, les introduire en input dans le modèle ALM de Cardif et de calculer le BEL pour les diverses fluctuations possibles.

Une étude préalable a contribué à l'analyse théorique des relations entre les variables et le BEL et à l'énoncé des liens attendus de l'analyse des données réelles. L'examen de ces dernières a, par la suite, permis de confirmer les prévisions de l'analyse théorique.

La base de données ainsi générée a servi pour calibrer divers modèles de Machine Learning, des plus simples comme la régression linéaire en passant par les GLM, aux plus complexes comme quelques algorithmes de Decision Trees (Adaboost, Random Forest, Gradient Boosting).

Les modèles ont tous présenté une performance correcte, certains même exceptionnelle expliquée par l'homogénéité des données. Autant que possible, les différences de performance ont été justifiées de manière théorique. Les résultats sur la base test et ceux du Backtesting ont permis de valider la fiabilité du modèle choisi : celui du Gradient Boosting.

L'approche choisie de choquer les variables, de les utiliser comme inputs au modèle ALM de BNP Paribas Cardif et d'en générer un BEL s'est avérée être, bien que naturelle en actuariat, assez coûteuse en temps et en efforts. Produire un grand nombre de données est presque incompatible avec Prophet, le logiciel sur lequel est implémenté le modèle ALM. Toutefois, au vu de l'émergence des méthodes de Machine Learning et donc de la nécessité de création de grandes données, Cardif s'oriente dans l'avenir vers Python pour la modélisation.

Par ailleurs, les choix effectués dans cette étude ne sont pas forcément optimaux et un

meilleur proxy pourrait naître d'une autre approche. Enfin, il serait intéressant de dupliquer une étude similaire afin d'élaborer des proxies d'autres indicateurs tels que le SCR, par exemple.

Bibliographie

AGBAHOLOU, T. (2019), '*3 cas d'analyse des scénarios économiques utilisés pour le calcul d'un Best Estimate en Epargne*. Institut des Actuaires'.
URL:
https ://www.institutdesactuaires.com/docs/mem/6be411c77bc590b70ba87c67748bfb5e.pdf

Best Estimate Liabilities Vie. Groupe de travail, Institut des Actuaires (2016).
URL:
https ://www.institutdesactuaires.com/global/gene/link.phpnews_link=2016110706_2016133822-npa41.pdf&fg=1

GERONDEAU, E. (2017), '*Ratio de couverture Solvabilité 2 d'un contrat d'épargne en euros, quels leviers de pilotage pour l'assureur ?*. Institut des Actuaires'.
URL:
https ://www.institutdesactuaires.com/docs/mem/3bba19a8cd26f7c5fc133d41bad3a095.pdf

HOUSSEINI, M. A. (2016), '*Anticipation de la déviation du BEL suivant différents états du monde*. Institut des Actuaires'.
URL:
https ://www.institutdesactuaires.com/docs/mem/bdb59c82655256c691b7e572dbd0b788.pdf

HULL, J. (2014), '*Options, Futures and Other Derivatives*. John Wiley & Sons'.

Information du souscripteur et du bénéficiaire. Code des Assurances (2017).
URL: *https ://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000035514611/*

JAMENI, A. (2019), '*Etude de la TVOG d'un contrat d'épargne euro sous IFRS 17*. Institut des Actuaires'.
URL:
https ://www.institutdesactuaires.com/docs/mem/8e748ce7a0e72a265bcfe0fbdfbc8c6e.pdf

MARBACH, B. (2013), '*Détermination et impact du coût du capital dans le cadre d'une étude d'allocation stratégique d'actifs : Exemple du Fonds Général de Cardif Assurance*

Vie. Institut des Actuaires'.
URL:
https ://www.institutdesactuaires.com/docs/mem/b5e6b82b59b48a98549341e248cf36e5.pdf

MONFERRINI, J. (2018), '*Projection des taux négatifs sous la probabilité monde réel*. Institut des Actuaires'.
URL:
https ://www.institutdesactuaires.com/docs/mem/2837a8da2ad71386b792bb39f95fcf4b.pdf

PIASER, P. (2013), '*Impact de la stratégie financière sur la PVFP et le capital réglementaire*. Ressources actuarielles'.
URL: *http ://www.ressources-actuarielles.net/EXT/ISFA/1226-02.nsf/0/d04f573bf37ac15ec1257cfd004d3bee/$FILE/me%CC%81moire%20-%20Pierrick%20Piaser.002.pdf/me%CC%81moire%20-%20Pierrick%20Piaser.pdf*

THEROND, P. (2013), '*Market consistency : fondements et limites*. Cours à l'ISFA, Université Lyon 1'.
URL: *http ://www.therond.fr/wpcontent/uploads/cours/MarketConsistency_2013.pdf*

VILA, G. (2019), '*Analyse du mouvement de best estimate épargne euro, basée sur la gestion fincière de l'assureur*. Institut des Actuaires'.
URL:
https ://www.institutdesactuaires.com/docs/mem/e8c40f83445fe91ea546690ac6e4be6b.pdf

Annexes

Annexe A

Prérequis de Finance

A.1 Les taux d'intérêt

A.1.1 L'actualisation

Afin de bien comprendre les taux zéro-coupon, il est nécessaire de comprendre le rôle des taux d'intérêt en général. En finance, on considère qu'un même actif n'a pas la même valeur au temps 0 et au temps $t > 0$ car on considère que :

- posséder avec certitude un actif aujourd'hui est moins risqué et donc plus valorisé que posséder le même actif dans le futur ;
- posséder un actif aujourd'hui présente divers avantages dans le sens où il peut être vendu et générer des gains ou placé à un taux avantageux (au minimum le taux sans risque), auquel cas il aura une valeur supérieure au temps $t > 0$.

Ce ne sont que certaines raisons intuitives parmi d'autres.

Prenons désormais un capital N (nominal) à titre d'exemple et supposons que l'on le place au taux sans risque $r > 0$ au temps t. Si l'on note donc $P(0,T)$ le prix aujourd'hui du nominal et que l'on observe son évolution jusqu'en T, on a :

$$P(0,T).(1+r) = P(T,T) \Leftrightarrow P(0,T) = \frac{P(T,T)}{1+r}$$

L'équation de gauche est une capitalisation (retrouver la valeur future à partir de la valeur actuelle) alors que l'équation de droite est une actualisation (retrouver la valeur actuelle à partir de la valeur future).

A.1.2 La fréquence de composition

Les taux d'actualisation sont généralement annuels. Cependant, il est important de faire attention à la composition de ce taux. En effet, un taux annuel de 10% composé semestriellement ou trimestriellement ne revient pas au même. A chaque composition, les intérêts générés par la composition précédente sont aussi capitalisés et à un taux différent.

A titre d'exemple, un nominal capitalisé au taux r semestriellement donnera $P(0,T).(1+\frac{r}{2})^2 = P(T,T)$ alors qu'un nominal capitalisé au taux r trimestriellement donnera $P(0,T).(1+\frac{r}{4})^4 = P(T,T)$. Plus généralement, pour un taux défini sur n unités de temps (une année, 5 mois ou 10 ans), à la fin de ces n unités de temps et si la fréquence de la composition est de m fois par unité de temps on a :

$$P(0,T).(1+\frac{r}{m})^{mn} = P(T,T)$$

A.1.3 La composition continue

Il existe une autre convention en Finance selon laquelle la composition se fait à une très haute fréquence. Mathématiquement cela se traduit par :

$$\lim_{m\to\infty} P(0,T).(1+\frac{r}{m})^{mn} = \lim_{m\to\infty} P(0,T)e^{mn.ln(1+\frac{r}{m})}$$

On se souvient du développement limité suivant au voisinage de 0 pour x :

$$ln(1+x) = x + o(x)$$

Pour m assez grand, $\frac{r}{m}$ est au voisinage de 0. D'où :

$$\begin{aligned} P(0,T).(1+\frac{r}{m})^{mn} &= P(0,T)e^{mn.(\frac{r}{m}+o(\frac{r}{m}))} \\ &= P(0,T)e^{nr(1+o(1))} \xrightarrow{m\to\infty} P(0,T)e^{nr} \end{aligned}$$

C'est le terme e^{nr} résultant du calcul qui permet la capitalisation continue. Une capitalisation continue (c'est-à-dire à fréquence infinie) d'un nominal $P(0,T)$ après n unités de temps donne $P(0,T)e^{nr}$. Réciproquement, un montant futur $P(T,T)$ vaut aujourd'hui $P(T,T)e^{-nr}$.

A.2 Les obligations

A.2.1 La valorisation d'une obligation

Une obligation est un actif qui permet à l'émetteur d'emprunter un nominal à l'acheteur en échange du remboursement de ce nominal en fin de vie de l'obligation (maturité) et parfois du paiement d'intérêts (uniques ou réguliers) appelés coupons jusqu'à l'arrivée à maturité.

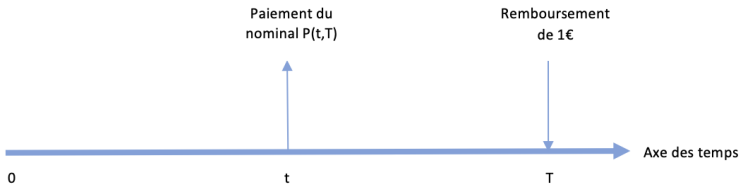

FIGURE A.1 – Le principe de l'actualisation au taux zéro-coupon

Une obligation zéro-coupon est une obligation dont l'achat ne permet pas de bénéficier de coupons.

Valoriser une obligation c'est connaitre son prix au temps 0, temps d'observation. Ceci revient à actualiser tous les flux que permet de toucher l'obligation dans l'avenir.

A.2.2 Le coupon couru

A un temps t fixé, durant la durée de vie d'une obligation, situé entre deux temps de paiements (ou détachement) de coupons i et $i+1$ ($i < t < i+1$), il n'y a pas de paiement de coupon. Le prochain coupon est toutefois à venir et il y a un cumul latent d'un pourcentage de ce coupon jusqu'au jour de détachement de ce dernier. Ce montant est le coupon couru. Pratiquement, il se calcule ainsi :

$$CC_t = \frac{t-i}{i+1-i}.Coupon_{i+1} = (t-i).Coupon_{i+1}$$

où :

- CC_t est le coupon couru au temps t intercalaire ;
- $Coupon_{i+1}$ est le coupon qui sera payé au temps $i+1$.

Cette notion de coupon couru sert à valoriser plus justement une obligation à un temps quelconque. Si cette dernière est rachetée à un temps t, la fraction du coupon couru à venir

cumulée avant cette date sera payée au nouveau détenteur de l'obligation mais doit être remboursée par ce dernier à l'ancien détenteur en guise de dédommagement. En effet, ces gains ont été cumulés pendant la période de détention de ce dernier [1]. Ainsi naît la cotation pied de coupon.

La cotation pied de coupon au temps t est simplement le prix de l'obligation à ce temps auquel est soustrait le coupon couru au temps t.

A.2.3 Les taux zéro-coupon

Le taux zéro-coupon (ou taux spot) est le taux de base qui permet d'actualiser les flux des obligations afin d'en connaitre le prix. Un taux zéro-coupon de maturité t ne permet, théoriquement, de n'actualiser qu'un flux qui a lieu au temps t [2]. Valoriser une obligation revient donc à actualiser tous les flux avec les taux correspondant repectivement aux maturités de ces flux. Ce taux est appelé zéro-coupon car il conrrespondrait au taux d'actualisation d'une obligation qui ne donne droit aux intérêts (en plus du nominal) qu'à maturité : c'est une obligation avec zéro-coupons intermédiaires.

Par exemple, soit une obligation de nominal 100€, de maturité 3 ans qui paie des coupons semestriellement au taux annuel de 4%. Soit $ZCR(t)$ le taux zéro-coupon associé à la maturité t.

Un taux annuel de 4% est un taux semestriel de $\frac{4}{2} = 2\%$. Chaque coupon vaut alors $0.02 * 100 = 2$€.

Le prix P_0 de cette obligation (avec une composition continue) est alors :

$$P_0 = 2e^{-0.5ZCR(0.5)} + 2e^{-1ZCR(1)} + 2e^{-1.5ZCR(1.5)} + 2e^{-2ZCR(2)}$$
$$+2e^{-2.5ZCR(2.5)} + (2 + 100)e^{-3ZCR(3)}$$

A.2.4 Le calcul des taux zéro-coupon

La courbe des taux zéro-coupon est donc la référence servant à valoriser les obligations. Pratiquement, cette courbe est calculée par bootstrapping à partir d'observations sur les

(1). Si l'achat se fait à la date $i+1$ du détachement d'un coupon ($t = i+1$), on considère que le coupon revient en totalité à l'ancien détenteur.

(2). En pratique et afin de simplifier, les traders utilisent parfois le même taux afin d'actualiser tous les flux. Cette approche est néanmoins considérée comme étant moins pertinente.

obligations d'Etat.

Les Etats sont toujours considérés comme solvables et les taux zéro-coupon des obligations qu'ils émettent sont ainsi considérés sans risque. Il est important de noter que le risque dont il est question ici est le risque de crédit (ou contrepartie). Il ne faut donc pas confondre ce taux sans risque (de crédit) et le taux r_t sans risque (de marché) qui concerne les placements court-terme ou liquidités.

Annexe B

Prérequis de Statistique

B.1 Test de Kolmogorov

Le test de Kolmogorov sert à vérifier l'hypothèse selon laquelle un échantillon $X_1, ..., X_n$ suit une loi donnée de fonction de répartition F. Afin d'effectuer ceci, le test procède par comparaison de fonctions de répartition.

Le test trie l'échantillon par ordre croissant (statistiques d'ordre) et définit la fonction dite de répartition empirique $\hat{F}$ comme suit :

$$\hat{F}(x) := \begin{cases} 0 & \text{si } x < X_1 \\ \frac{i}{n} & \text{si } X_i \leq x < X_{i+1} \\ 1 & \text{si } x \geq X_n \end{cases}$$

Est calculée ensuite la distance suivante :

$$D_K(\hat{F}, F) := \max_{i=1,...,n} \left\{ \left| F(X_i) - \frac{i}{n} \right|, \left| F(X_i) - \frac{i-1}{n} \right| \right\}$$

Il suffit alors de comparer la distance $D_K(\hat{F}, F)$ par rapport aux valeurs tabulées du test. Si $D_K(\hat{F}, F)$ est supérieur alors il est possible de rejeter l'hypothèse.

B.2 Modèles de Machine Learning basés sur les arbres de décision

B.2.1 Decision Trees

Un *arbre de décision* est un outil d'aide à la décision qui, à l'issue d'une série de questions, offre une suggestion de décision à l'utilisateur. En apprentissage, les choix sont effectués selon des réponses à des questions sur les données. L'aboutissement de l'algorithme fournit une régression [(1)] de la variable en question.

B.2.2 Adaboost

AdaBoost est un algorithme d'apprentissage statistique appartenant à la famille des algorithmes de *Boosting* i.e les algorithmes qui reposent sur la réduction itérative de l'erreur des machines "faibles" qui ont un rendement à peine supérieur à un algorithme retournant un résultat au hasard.

AdaBoost fut initialement créée pour des tâches de classification et ensuite adaptée pour des tâches de régression (tâche pour laquelle il est utilisé dans ce mémoire).

AdaBoost.R2, qui est l'algorithme codé dans scikit learn et utilisé dans ce contexte, est une modification de AdaBoost.R et dont l'idée est de commencer par donner des poids w_i égaux à chaque observation X_i, de définir ensuite un estimateur dont on calcule l'erreur à travers une fonction de coût L puis d'actualiser les poids w_i selon l'erreur du précédent estimateur. L'algorithme (original) s'arrête quand la fonction de coût est inférieure ou égale à 0.5.

B.2.3 Random Forest

Random Forest Regression est un algorithme d'apprentissage statistique supervisé qui permet de combiner plusieurs algorithmes d'apprentissage (notamment des arbres de décisions) afin d'avoir une meilleure précision.

L'algorithme fonctionne suivant la technique de *Bagging* : il détermine le nombre d'arbres dans la forêt. A chaque étape, il génère un vecteur $\Theta_k = (X_k, Y_K)$ (tous de même loi et indépendants) aléatoirement parmi les données. En utilisant ce vecteur Θ_k, il construit un

(1). Les arbres de décision peuvent aussi s'appliquer aux problèmes de classification mais il ne s'agit pas du cadre de cette étude car le BEL est une variable continue.

estimateur (arbre de décision). Ainsi, le résultat pour une nouvelle donnée est la moyenne des estimations pour chacun des N estimateurs construits précédemment.

Annexe C

Code Python

```
#Preprocessing

'''I- Importing packages'''

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.api as sm
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn import metrics
from sklearn import tree
from sklearn.ensemble import AdaBoostRegressor
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn import tree
import seaborn as sns
sns.set()
import warnings
warnings.filterwarnings('ignore')
np.random.seed(1234)
```

```
'''II- Import des données'''

data = pd.read_csv("BDD.csv", sep=";")
data = pd.DataFrame(data)

'''III- On efface les lignes à BEL nul, car ces lignes manquent de données et
on construit X et Y'''

data = data[data["BEL"]!=0]
X1, X2, X3, X4, X5, X6, X7, X8, X9, X10=(data[["PMVL"]], data[["PPE"]],
                                         data[["BETA_0"]], data[["BETA_1"]],
                                         data[["BETA_2"]], data[["LAMBDA"]],
                                         data[["PDD"]], data[["PM"]],
                                         data[["RC"]], data[["CASH"]])

#-------------------------------------------------------------------------------
#Data visualisation

'''I- Résumé, boxplot et distribution'''

(data[["BEL"]]).describe()
(data[["BEL"]]).boxplot()

sns.distplot(data[["BEL"]])
plt.xlabel("BEL")
plt.ylabel("Densité")
plt.title("Distribution du BEL")

'''II- BEL selon quelques variables'''

'''1) BEL selon PMVL'''

data_pmvl=data[((data["PPE"]==data.iloc[2][3]) & (data["BETA_0"]==data.iloc[1][4]) &
                (data["BETA_1"]==data.iloc[1][5]) & (data["BETA_2"]==data.iloc[1][6])&
                (data["LAMBDA"]==data.iloc[1][7]))]
plt.scatter(data_pmvl[["PMVL"]], data_pmvl[["BEL"]], label="PMVL")
```

```
plt.legend()
plt.xlabel("PMVL")
plt.ylabel("BEL")
plt.title("BEL selon la PMVL")

'''2) BEL selon PPE'''

data_ppe=data[((data["PMVL"]==data.iloc[0][2]) & (data["BETA_0"]==data.iloc[1][4]) &
               (data["BETA_1"]==data.iloc[1][5]) & (data["BETA_2"]==data.iloc[1][6]) &
               (data["LAMBDA"]==data.iloc[1][7]))]
plt.scatter(data_ppe[["PPE"]], data_ppe[["BEL"]], label="PPE")
plt.legend()
plt.xlabel("PPE")
plt.ylabel("BEL")
plt.title("BEL selon la PPE")

#########pour avoir la pmvl et la ppe sur un même graph
plt.legend()
plt.xlabel("Variable")
plt.ylabel("BEL")
plt.title("BEL selon la PMVL et la PPE")

data_nelson=data[(data["PMVL"]==data.iloc[0][2]) & (data["PPE"]==data.iloc[2][3])]

plt.scatter(data_nelson[["BETA_0"]], data_nelson[["BEL"]], label=r'$\beta_0$')
plt.scatter(data_nelson[["BETA_1"]], data_nelson[["BEL"]], label=r'$\beta_1$')
plt.scatter(data_nelson[["BETA_2"]], data_nelson[["BEL"]], label=r'$\beta_2$')
plt.scatter(data_nelson[["LAMBDA"]], data_nelson[["BEL"]], label=r'$\lambda$')
plt.legend()
plt.xlabel("Paramètres Nelson-Siegel")
plt.ylabel("BEL")
plt.title("BEL selon les indicateurs Nelson-Siegel")

data.drop('ANNEE', inplace=True, axis=1)

'Brassage des données'
```

```
data=data.sample(frac=1)

'Division backtesting et apprentissage'

data_backtesting=data[data["ANNEE"]==2020]
data=data[data["ANNEE"]!=2020]

#Correlation heat map

'''On essaie de traduire la corrélation entre les variables'''

'''I- Corrélation Pearson'''

X = data[["PMVL", "PPE", "BETA_0", "BETA_1", "BETA_2", "LAMBDA", "PDD", "PM",
          "RC", "CASH"]]
Y = data[["BEL"]]

X_backtesting = data_backtesting[["PMVL", "PPE", "BETA_0", "BETA_1", "BETA_2",
                                  "LAMBDA", "PDD", "PM", "RC", "CASH"]]
Y_backtesting = data_backtesting[["BEL"]]

matrix1 = X.corr(method='pearson')
ax = sns.heatmap(
    matrix1,
    vmin = -1, vmax = 1, center = 0,
    cmap = sns.diverging_palette(20, 220, n = 200),
    square = True
)
ax.set_xticklabels(
    ax.get_xticklabels(),
    rotation = 45,
    horizontalalignment = 'right'
);

ax.set_title("Matrice de corrélation Pearson")

'''II Corrélation Spearman'''
```

```
matrix2 = X.corr(method='spearman')
ax = sns.heatmap(
    matrix1,
    vmin = -1, vmax = 1, center = 0,
    cmap = sns.diverging_palette(20, 220, n = 200),
    square = True
)
ax.set_xticklabels(
    ax.get_xticklabels(),
    rotation = 45,
    horizontalalignment = 'right'
);

ax.set_title("Matrice de corrélation Spearman")

X.drop('RC', inplace=True, axis=1)
X.drop('CASH', inplace=True, axis=1)
data=pd.concat([X, Y], axis=1)

X_backtesting.drop('RC', inplace=True, axis=1)
X_backtesting.drop('CASH', inplace=True, axis=1)

#ACP

'''ACP'''

sc = StandardScaler()
data_ACP = sc.fit_transform(X)

pca = PCA().fit(data_ACP)
r = pca.explained_variance_ratio_.cumsum()

plt.plot(r)
plt.title("Ratio of explained variance")

plt.plot(4, r[4], 'o')
```

```
plt.annotate("{0:0.2f}".format(r[4]), xy=(4, r[4]),
             xytext=np.r_[4, r[4]]+np.r_[0.1, -0.03])

plt.plot(6, r[6], 'o')
plt.annotate("{0:0.2f}".format(r[6]), xy=(6, r[6]),
             xytext=np.r_[6, r[6]]+np.r_[0.1, -0.03])

plt.xlabel("Nombre de variables")
plt.ylabel("Ratio")

pca1 = PCA(n_components=2).fit(data_ACP)

def scatterc(X, labels=None):
    plt.scatter(X[:, 0], X[:, 1], c=np.arange(X.shape[0]))
    if labels is not None:
        for x, label in zip(X, labels):
            plt.annotate(label, xy=x,
                         xytext=x + np.r_[0.02, 0.02]*(X.max(axis=0)-X.min(axis=0)))

scatterc(pca1.components_.T, X.columns)
plt.xlabel("Premier axe")
plt.ylabel("Deuxième axe")
plt.title("Corrélation des variables selon les 2 premiers axes principaux")

#Base de données train et test

'''70 et 30%'''

X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.3)

#-------------------------------------------------------------------------------
#Apprentissage supervisé sans validation croisée

x=np.linspace(90000000000,123000000000,100)
y=x
scores_train={}
scores_test={}
```

```
scores_cross={}

#----------------------------Classic regressions: linear and GLM

'''I- Régressions classiques'''

'''1) Régression linéaire'''

'''i) Construction du modèle'''

regression = LinearRegression().fit(X_train, Y_train)
beta1, beta0=regression.coef_, regression.intercept_
regression_score = regression.score(X_train, Y_train)
print(regression_score)

'''ii) Prédictions'''

Y_pred_reg=regression.predict(X_test)
regression_score_pred = regression.score(X_test, Y_test)
print(regression_score_pred)
print('Mean Absolute Error:', metrics.mean_absolute_error(Y_test, Y_pred_reg))
print('Mean Squared Error:', metrics.mean_squared_error(Y_test, Y_pred_reg))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(Y_test,
                                                                    Y_pred_reg)))
print(("Erreur en pourcentage par rapport à un BEL moyen:",
       np.sqrt(metrics.mean_squared_error(Y_test, Y_pred_reg))*100/Y_test.mean()))
plt.scatter(Y_test, Y_pred_reg)
plt.plot(x, y, c='r')
plt.axis('off')
plt.xlabel("Y_test")
plt.ylabel("Y_pred")
plt.title("Régression Linéaire : adéquation entre les $Y_{test_i}$ et les $\widehat{Y}_{

'''iii) BIC et erreurs'''

'BIC=n*LL+k*log(n)'
```

```
def BIC_reg(n, mse, num_param):
    return n * np.log(mse) + num_param * np.log(n)

mse=metrics.mean_squared_error(Y_test, Y_pred_reg)
num_param=len(regression.coef_)+1
n=len(Y_test)
BIC_reg=BIC_reg(n, mse, num_param)

erreurs_regression=np.sqrt(metrics.mean_squared_error(Y_test, Y_pred_reg))*100/Y_test
erreurs_regression.boxplot()

'''iv) Validation croisée'''

cross_score_reg = np.mean(cross_val_score(regression, X, Y, cv=5))
print(cross_score_reg)

def with_intercept(X):
    N = len(X)
    p = len(X.columns) + 1
    X_with_intercept = np.empty(shape=(N, p), dtype=np.float)
    X_with_intercept[:, 0] = 1
    X_with_intercept[:, 1:p] = X.values
    return X_with_intercept

'''2) Modèle Linéaire Généralisé (GLM)'''

'''a) Gamma'''

'''i) Construction du modèle et summary'''
gamma_model = sm.GLM(Y_train, with_intercept(X_train), family=sm.families.Gamma())
gamma_results = gamma_model.fit()
print(gamma_results.summary())

'''ii) Prédictions'''

Y_pred_gamma=gamma_results.predict(with_intercept(X_test))
```

```
print('Mean Absolute Error:', metrics.mean_absolute_error(Y_test, Y_pred_gamma))
print('Mean Squared Error:', metrics.mean_squared_error(Y_test, Y_pred_gamma))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(Y_test,
                                                                    Y_pred_gamma)))
print(("Erreur en pourcentage par rapport à un BEL moyen:",
       np.sqrt(metrics.mean_squared_error(Y_test, Y_pred_gamma))*100/Y_test.mean()))
plt.scatter(Y_test, Y_pred_gamma)
plt.plot(x, y, c='r')
plt.axis('off')
plt.xlabel("Y_test")
plt.ylabel("Y_pred")
plt.title("Gamma GLM : adéquation entre les $Y_{test_i}$ et les $\widehat{Y}_{test_i}$")

'''iii) BIC et erreurs'''

BIC_gamma=gamma_results.bic

erreurs_gamma=np.sqrt(metrics.mean_squared_error(Y_test, Y_pred_gamma))*100/Y_test
erreurs_gamma.boxplot()

'''b) Gaussian'''

'''i) Construction du modèle et summary'''
gaussian_model = sm.GLM(Y_train, with_intercept(X_train),
                        family=sm.families.Gaussian())
gaussian_results = gaussian_model.fit()
print(gaussian_results.summary())

'''ii) Prédictions'''

Y_pred_gaussian=gaussian_results.predict(with_intercept(X_test))
print('Mean Absolute Error:', metrics.mean_absolute_error(Y_test, Y_pred_gaussian))
print('Mean Squared Error:', metrics.mean_squared_error(Y_test, Y_pred_gaussian))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(Y_test,
                                                                    Y_pred_gaussian)))
print(("Erreur en pourcentage par rapport à un BEL moyen:",
       np.sqrt(metrics.mean_squared_error(Y_test, Y_pred_gaussian))*100/Y_test.mean()))
```

```
plt.scatter(Y_test, Y_pred_gaussian)
plt.plot(x, y, c='r')
plt.axis('off')
plt.xlabel("Y_test")
plt.ylabel("Y_pred")
plt.title("Gaussian GLM : adéquation entre les $Y_{test_i}$ et les $\widehat{Y}_{test_i}

'''iii) BIC et erreurs'''

BIC_gaussian=gaussian_results.bic

erreurs_gaussian=np.sqrt(metrics.mean_squared_error(Y_test,
                                                    Y_pred_gaussian))*100/Y_test
erreurs_gaussian.boxplot()

#----------------------------Trees

'''I- Decision Trees'''

'''i) Construction du modèle'''

decision_tree = tree.DecisionTreeRegressor().fit(X_train, Y_train)
decision_tree_score = decision_tree.score(X_train, Y_train)
print(decision_tree_score)

'''ii) Prédictions'''

Y_pred_tree=decision_tree.predict(X_test)
decision_tree_score_pred = decision_tree.score(X_test, Y_test)
print(decision_tree_score_pred)
print('Mean Absolute Error:', metrics.mean_absolute_error(Y_test, Y_pred_tree))
print('Mean Squared Error:', metrics.mean_squared_error(Y_test, Y_pred_tree))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(Y_test,
                                                                    Y_pred_tree)))
print(("Erreur en pourcentage par rapport à un BEL moyen:",
       np.sqrt(metrics.mean_squared_error(Y_test, Y_pred_tree))*100/Y_test.mean()))
plt.scatter(Y_test, Y_pred_tree)
```

```
plt.plot(x, y, c='r')
plt.xlabel("Y_test")
plt.ylabel("Y_pred")
plt.title("Decision Trees : adéquation entre les $Y_{test_i}$ et les $\widehat{Y}_{test_

'''iii) Validation croisée'''

cross_score_tree = np.mean(cross_val_score(decision_tree, X, Y, cv=5))
print(cross_score_tree)

'''iv) Erreurs'''

erreurs_tree=np.sqrt(metrics.mean_squared_error(Y_test, Y_pred_tree))*100/Y_test
erreurs_tree.boxplot()

'''II- Adaboost'''

'''i) Construction du modèle'''

adaboost = AdaBoostRegressor().fit(X_train, Y_train)
adaboost_score = adaboost.score(X_train, Y_train)
print(adaboost_score)

'''ii) Prédictions'''

Y_pred_adaboost=adaboost.predict(X_test)
adaboost_score_pred = adaboost.score(X_test, Y_test)
print(adaboost_score_pred)
print('Mean Absolute Error:', metrics.mean_absolute_error(Y_test, Y_pred_adaboost))
print('Mean Squared Error:', metrics.mean_squared_error(Y_test, Y_pred_adaboost))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(Y_test,
                                                                    Y_pred_adaboost)))
print(("Erreur en pourcentage par rapport à un BEL moyen:",
       np.sqrt(metrics.mean_squared_error(Y_test, Y_pred_adaboost))*100/Y_test.mean()))
plt.scatter(Y_test, Y_pred_adaboost)
plt.plot(x, y, c='r')
plt.xlabel("Y_test")
```

```
plt.ylabel("Y_pred")
plt.title("Adaboost : adéquation entre les $Y_{test_i}$ et les $\widehat{Y}_{test_i}$")

'''iii) Validation croisée'''

cross_score_adaboost = np.mean(cross_val_score(adaboost, X, Y, cv=5))
print(cross_score_adaboost)

'''iv) Erreurs'''

erreurs_adaboost=np.sqrt(metrics.mean_squared_error(Y_test,
                                                    Y_pred_adaboost))*100/Y_test
erreurs_adaboost.boxplot()

'''III- Gradient Boosting'''

'''i) Construction du modèle'''

gradient = GradientBoostingRegressor().fit(X_train, Y_train)
gradient_score = gradient.score(X_train, Y_train)
print(gradient_score)

'''ii) Prédictions'''

Y_pred_gradient=gradient.predict(X_test)
gradient_score_pred = gradient.score(X_test, Y_test)
print(gradient_score_pred)
print('Mean Absolute Error:', metrics.mean_absolute_error(Y_test, Y_pred_gradient))
print('Mean Squared Error:', metrics.mean_squared_error(Y_test, Y_pred_gradient))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(Y_test,
                                                                    Y_pred_gradient)))
print(("Erreur en pourcentage par rapport à un BEL moyen:",
       np.sqrt(metrics.mean_squared_error(Y_test, Y_pred_gradient))*100/Y_test.mean()))
plt.scatter(Y_test, Y_pred_gradient)
plt.plot(x, y, c='r')
plt.axis('off')
plt.xlabel(r'$Y_i$')
```

```
plt.ylabel(r'$\widehat{Y_i}$')
plt.title('Gradient Boosting : adéquation entre les $Y_{test_i}$ et les $\widehat{Y}_{te

'''iii) Validation croisée'''

cross_score_gradient = np.mean(cross_val_score(gradient, X, Y, cv=5))
print(cross_score_gradient)

'''iv) Erreurs'''

erreurs_gradient=np.sqrt(metrics.mean_squared_error(Y_test,
                                                    Y_pred_gradient))*100/Y_test
erreurs_gradient.boxplot()

'''IV- Random Forest'''

'''i) Construction du modèle'''

forest = RandomForestRegressor().fit(X_train, Y_train)
forest_score = forest.score(X_train, Y_train)
print(forest_score)

'''ii) Prédictions'''

Y_pred_forest=forest.predict(X_test)
forest_score_pred = forest.score(X_test, Y_test)
print(forest_score_pred)
print('Mean Absolute Error:', metrics.mean_absolute_error(Y_test, Y_pred_forest))
print('Mean Squared Error:', metrics.mean_squared_error(Y_test, Y_pred_forest))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(Y_test,
                                                                    Y_pred_forest)))
print(("Erreur en pourcentage par rapport à un BEL moyen:",
       np.sqrt(metrics.mean_squared_error(Y_test, Y_pred_forest))*100/Y_test.mean()))
plt.scatter(Y_test, Y_pred_forest)
plt.plot(x, y, c='r')
plt.xlabel("Y_test")
plt.ylabel("Y_pred")
```

```
plt.title("Random Forest : adéquation entre les $Y_{test_i}$ et les $\widehat{Y}_{test_i

'''iii) Validation croisée'''

cross_score_forest = np.mean(cross_val_score(forest, X, Y, cv=5))
print(cross_score_forest)

'''iv) Erreurs'''

erreurs_forest=np.sqrt(metrics.mean_squared_error(Y_test, Y_pred_forest))*100/Y_test
erreurs_forest.boxplot()

#----------------------------Résumé des modèles

erreurs=[]
erreurs = pd.concat([erreurs_regression, erreurs_gamma, erreurs_gaussian,
                     erreurs_tree, erreurs_adaboost, erreurs_gradient,
                     erreurs_forest], axis=1)
erreurs.columns=['Régression Linéaire', 'GLM Gamma', 'GLM Gassian',
                 'Decision Tree', 'Adaboost', 'Gradient Boosting', 'Forêt Aléatoire']
erreurs.boxplot(figsize=(20, 13), fontsize=15)

#----------------------------Backtesting


Y_pred_backtesting=gradient.predict(X_backtesting)

print('Mean Absolute Error:', metrics.mean_absolute_error(Y_backtesting,
                                                          Y_pred_backtesting))
print('Mean Squared Error:', metrics.mean_squared_error(Y_backtesting,
                                                        Y_pred_backtesting))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(Y_backtesting,
                                                                    Y_pred_backtesting)
print(("Erreur en pourcentage par rapport à un BEL moyen:",
       np.sqrt(metrics.mean_squared_error(Y_backtesting,
                                          Y_pred_backtesting))*100/Y_backtesting.mean())
```

```
plt.scatter(Y_backtesting, Y_pred_backtesting)
plt.plot(x, y, c='r')
plt.axis('off')
plt.xlabel(r'$Y_i$')
plt.ylabel(r'$\widehat{Y_i}$')
plt.title('Gradient Boosting : adéquation entre les $Y_{i}$ et les $\widehat{Y}_{i}$')
```

Annexe D

Code R

Les premières lignes du code R (qui constituent l'input de la courbe des taux) sont omises pour des raisons de confidentialité. Elles ont permis de définir les vecteurs *taux_central_18*, *taux_10_18*, *taux_m10_18*, *taux_central_19*, *taux_10_19*, *taux_m10_19*, *taux_central_20*, *taux_10_20* et *taux_m10_20*.

```
library(YieldCurve)

maturite=c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 30)

model_central_18=Nelson.Siegel(taux_central_18, maturite)
model_10_18=Nelson.Siegel(taux_10_18, maturite)
model_m10_18=Nelson.Siegel(taux_m10_18, maturite)
model_central_19=Nelson.Siegel(taux_central_19, maturite)
model_10_19=Nelson.Siegel(taux_10_19, maturite)
model_m10_19=Nelson.Siegel(taux_m10_19, maturite)
model_central_20=Nelson.Siegel(taux_central_20, maturite)
model_10_20=Nelson.Siegel(taux_10_20, maturite)
model_m10_20=Nelson.Siegel(taux_m10_20, maturite)

z=c(0.02098761, 0.9632928, 1.190334, 0.1236869)
nelson=function(beta0, beta1, beta2, lambda, matu)
{return(beta0+(beta1*((1-exp(-matu*lambda))/(matu*lambda)))+
          (beta2*((((1-exp(-matu*lambda))/(matu*lambda)))-(exp(-matu*lambda)))))}
y=nelson(0.02098761, 0.9632928, 1.190334, 0.1236869, maturite)
plot(maturite, taux_central_18, col="red",
```

```
     main="Régression Nelson pour la courbe centrale 2018", pch=18,
     xlab="Maturité", ylab="Taux")
lines(maturite, y, col="blue")
legend("topright", legend=c("Taux", "Nelson"), col=c("red", "blue"), lty=c(0,1),
       pch = c(18, NA))
```