

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

Par : Monsieur Baptiste Voisin

Titre du mémoire : La valeur potentielle des clients professionnels

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

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

*Membres présents du jury de
l'Institut des Actuaire*

signature

Entreprise :

Nom :

Generali

Signature :

*Directeur de mémoire en
entreprise :*

*Membres présents du jury de la
filière*

Nom : Samia Nouar

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



Master 2 d'actuariat ISUP

La valeur potentielle des clients professionnels Generali - Valeur Client

Auteur :
Baptiste VOISIN

Directrice de mémoire en entreprise :
Samia NOUAR

Directrice de mémoire académique :
Claire BOYER



Confidentiel
10 décembre 2019

Résumé

Le marché très concurrentiel du secteur des assurances, notamment sur les produits d'appel, impose aux compagnies de cibler leurs clients à potentiel de façon plus fine qu'auparavant. Dans cette perspective, l'un des dispositifs mis en place à Generali a été le développement d'une vision transverse du client. Ainsi, deux représentations de la rentabilité coexistent désormais, l'une au niveau des produits et l'autre au niveau du client.

C'est dans ce cadre que les notions de valeur du client et de valeur potentielle ont été déployées. La valeur du client correspond à la projection actualisée des résultats techniques et financiers des contrats du client sur une période probable de présence en portefeuille, tandis que la valeur potentielle est la rentabilité probable des futurs contrats du client.

L'objectif de ce mémoire est le développement d'un modèle de valeur potentielle uniquement sur les clients professionnels. Actuellement, il n'existe qu'un seul modèle, mélangeant ainsi les clients particuliers et professionnels. Or, ces derniers présentent un potentiel beaucoup important. En effet, le volume de chiffre d'affaires et le nombre moyen de contrats sont plus élevés que pour les clients particuliers.

La modélisation de la valeur potentielle se fera, dans ce mémoire, en deux étapes. La première consiste en la modélisation de la probabilité que le client souscrive un nouveau contrat dans l'année. La seconde étape est la modélisation de la rentabilité du contrat en cas de souscription. Finalement, la valeur potentielle correspond au produit de la probabilité et de la rentabilité.

Enfin, une fois le modèle de valeur potentielle validé, une application sur l'optimisation des offres commerciales pour les prospects est proposée. Dans cette application, différents scénarios sont simulés, dans lesquels le niveau de rabais proposé varie en fonction de la valeur potentielle des prospects à potentiel. Les scénarios considérés diffèrent par le niveau de rabais et par le nombre de prospects à potentiel considérés. Le but de cette application est de voir comment varie la valeur totale du portefeuille de prospects en fonction des scénarios.

Abstract

The highly competitive insurance market, particularly for call products, requires companies to target their potential customers more precisely than before. With that in mind Thus, two representations of profitability now coexist, one at the product level and the other at the customer level. With this in mind, one of the measures implemented at Generali has been the development of a transversal vision of the customer. Thus, two representations of profitability now coexist, one at the product level and the other at the customer level.

It is in this context that the notions of customer value and potential value have been deployed. The customer value is the projection of the technical and financial results of the customer's contracts over a probable period of time in the portfolio, while the potential value is the probable profitability of the customer's future contracts.

The objective of this thesis is to develop a potential value model only for professional clients. Currently, there is only one model, mixing private and professional customers. However, the latter have a great deal of potential. Indeed, the volume of revenue and the average number of contracts are higher than for individual customers.

The modelling of the potential value will be done in two steps in this paper. The first is to model the probability that the client will take out a new contract within the year. The second step is to model the profitability of the contract in the event of subscription. Finally, the potential value is the product of probability and profitability.

Finally, once the potential value model has been validated, an application on optimizing sales offers for prospects is proposed. In this application, different scenarios are simulated, in which the proposed discount level varies according to the potential value of the potential prospects with potential. The scenarios considered differ in the level of discounts and the number of potential prospects considered. The purpose of this application is to see how the total value of the lead portfolio varies according to the scenarios.

Note de synthèse

Problématique

La valeur client est un indicateur de la rentabilité du client. Il est composé de la **valeur des contrats** et de la **valeur potentielle** du client.

- La valeur des contrats correspond à la projection des marges technique et financière de chacun des produits d'assurance.
- La valeur potentielle est la valeur des futurs contrats pondérée par la probabilité de souscription de ces contrat.

Actuellement le modèle de valeur potentielle est le même pour tous les clients de la compagnie. Ainsi il n'y a pas de différenciation nette entre les clients particuliers et les clients professionnels. L'objet de ce mémoire est de **développer un modèle uniquement autour des clients professionnels** dans le but de pouvoir identifier de nouveaux segments de clients à potentiel.

Le modèle qui va être développé est séparé en deux parties :

1. Tout d'abord une probabilité de multi-équipement est calculée via des méthodes d'apprentissage statistique.
2. Ensuite la distribution de la valeur en cas de multi-équipement est calculée.

Modélisation

Modéliser la probabilité de multi-équipement

Les données qui sont utilisées dans ce mémoire pour construire le modèle correspondent aux professionnels en portefeuille de 2014 à 2016. La base de données constituée des professionnels en portefeuille en 2017 est utilisée comme échantillon de test pour le modèle calculant la probabilité de multi-équipement. Cela permet de vérifier la stabilité des probabilités dans le temps.

La probabilité de multi-équipement est calculée via deux modèles mis en concurrence : une régression logistique et un gradient boosting.

1. Le modèle de régression logistique :

Le modèle de régression logistique est parcimonieux dans le mesure où peu de variables explicatives sont utilisées. Cette sélection de variable est réalisé via une pénalisation Lasso du modèle.

Le modèle retenu possède 20 variables explicatives.

Les variables qui poussent le plus la probabilité prédite vers de fortes valeurs sont :

- le fait de posséder au moins un contrat *Automobile* ;
- le fait de posséder au moins un contrat *Flotte Automobile*.

Les variables qui poussent le plus la probabilité prédite vers de faibles valeurs sont :

- le fait d'appartenir au réseau de distribution '*Coutiers ou assimilés*' qui correspond au réseau de courtiers et assimilés (hors Grands Courtiers) ;
- le fait d'être mono-équipé en produit *Vie* ou d'assurance de *Personnes* ;
- le fait d'avoir une ancienneté forte.

2. Le modèle de gradient boosting :

Ce type de modèle est en général plus efficace car il permet de mieux capter les interactions entre les différentes variables explicatives. Toutefois il a le désavantage d'être difficilement interprétable. De plus certains paramètres ne sont pas optimisés par l'algorithme, il s'agit des hyperparamètres. Deux d'entre eux, la profondeur et le nombre d'arbres, sont déterminés par validation croisée.

La statistique d'**importance des variables** est utilisée pour déterminer les variables explicatives apportant les gains d'information les plus élevés au modèle.

Les variables qui ont l'impact le plus important sont :

- (a) *le nombre de contrats Automobile* ;
- (b) *l'ancienneté du contrat le plus récent* ;
- (c) *l'ancienneté du client* ;
- (d) *le nombre de contrats détenus*.

3. Comparaison des modèles :

La principale statistique de performance du modèle qui est utilisée est l'aire sous la courbe ROC, elle est résumée dans le tableau ci-dessous :

Modèle	Base de données 2014 - 2016	Base de données 2017
Régression logistique	74 %	79 %
Gradient boosting	78 %	81 %

Les performances augmentent pour la base de données de 2017 car celle-ci présente une proportion d'individus se multi-équipant plus faible. Or la courbe ROC est un indicateur sensible aux bases déséquilibrées. Toutefois la performance constatée est meilleure dans les deux cas pour le modèle de gradient boosting.

Nous avons de plus choisi de représenter de façon graphique les performances afin de juger les modèles de manière qualitative. Sur cette représentation graphique, les prédictions

sont triées et regroupées par paquet de 3000. A sein de chaque regroupement la moyenne des prédictions et la moyenne des vraies sont calculées. Sur la figure 1, la courbe verte correspond à la moyenne des vraies valeurs et la courbe noire en escalier correspond à la moyenne des prédictions. Enfin la droite bleue correspond au taux moyen de multi-équipement sur toute la base de données.

Cette représentation graphique permet de mesurer la volatilité des prédictions par rapport aux valeurs réelles. Le fait d'avoir 3000 individus par groupe permet de réduire la volatilité dans le calcul des moyennes.

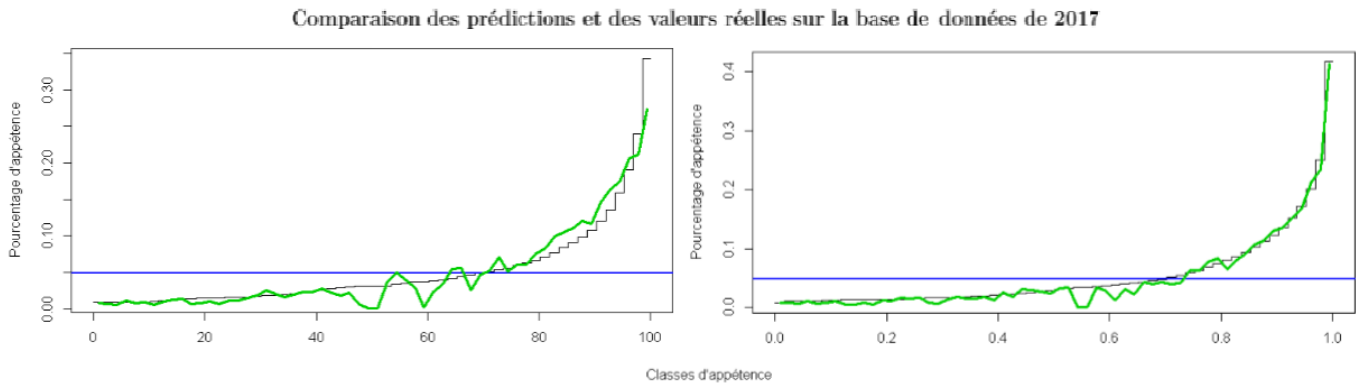


Figure 1 – À gauche : la régression logistique, à droite : le gradient boosting

Pour la suite du modèle, il a été décidé de garder le modèle de gradient boosting qui présente de meilleures performances.

Modéliser la rentabilité en cas de souscription

La rentabilité en cas de souscription est modélisée en tant que la valeur moyenne du produit souscrit multipliée par la probabilité d'équipement du dit produit :

$$\text{Valeur moyenne en cas de souscription} = \sum_{i \in \text{Produits}} p_i \times \bar{V}_i$$

- **Construction des p_i :**

La probabilité de souscription vers le produit i , sachant que le client se multi-équipe, correspond au pourcentage de clients s'équipant vers ce produit. De plus les clients sont divisés en groupes en fonction de leur nature (professionnel ou entreprise) et de leur panier de produits.

Les différents produits considérés sont :

L'Automobile hors les engins spéciaux, l'Habitation, les engins spéciaux, l'assurance des commerçants, l'épargne et la retraite, l'assurance de Personnes et les Autres produits IARD.

Ainsi nous aboutissons aux différentes probabilités résumées dans le tableau ci-dessous :

Classe du client	Probabilité de souscrire le produit :						
	Auto hors E.G.	Autre IARD	Engins Spéciaux	Epargne Retraite	MRC	MRH + PNO	Prévoyance
Non PRO - IARD 00	24%	26%	4%	1%	23%	19%	4%
Non PRO - IARD 01	12%	57%	9%	1%	13%	8%	1%
Non PRO - IARD 10	53%	18%	9%	1%	14%	4%	2%
Non PRO - IARD 11	38%	27%	19%	1%	9%	3%	2%
Non PRO - Mixte	39%	23%	9%	4%	13%	6%	6%
Non PRO - Assurance de Pers.	12%	14%	0%	10%	14%	2%	47%
Non PRO - Assurance Vie	4%	5%	0%	73%	7%	0%	12%
PRO - IARD 00	43%	13%	3%	1%	8%	18%	15%
PRO - IARD 01	29%	25%	15%	0%	4%	17%	10%
PRO - IARD 10	52%	9%	5%	0%	3%	23%	7%
PRO - IARD 11	43%	13%	12%	0%	4%	22%	6%
PRO - Mixte	36%	10%	4%	3%	3%	23%	21%
PRO - Assurance de Pers.	13%	5%	1%	5%	2%	11%	63%
PRO - Assurance Vie	10%	1%	0%	27%	1%	2%	60%

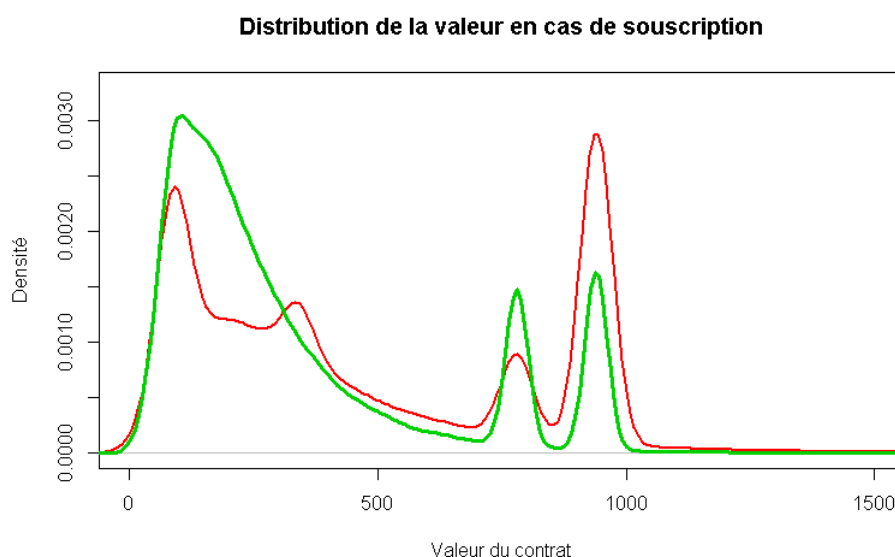
Figure 2 – Probabilité de souscription des produits équipés

• Construction de $\bar{V}_i$:

- Pour les produits présentant un volume de souscription peu important : les $\bar{V}_i$ correspondent à la moyenne de la valeur des affaires nouvelles réalisées sur le produit i .
- Pour les produits présentant un volume de souscription important (*l'Automobile hors les engins spéciaux, l'Habitation et les engins spéciaux*) : des sous-populations sont identifiées via des arbres de régression au sein desquelles la moyenne de la valeur des contrats est calculée.

La volatilité restant importante au sein des groupes déterminés par les arbres de régression, il a été décidé d'ajuster une loi de probabilité pour quantifier cet aléa.

Ainsi, grâce à l'ajustement de ces lois, nous disposons de la distribution de la rentabilité en cas de souscription en tant qu'une loi mélange des lois précédemment ajustée et de lois Dirac pour les produits peu souscrits. Les probabilités de souscription correspondant aux poids de chacune des lois.



Ci-dessus sont représentées les distributions des rentabilité en cas de souscription pour deux individus différents :

- L'individu rouge est une entreprise possédant des contrats *Vie* et *Non-Vie*.
- L'individu vert est un professionnel possédant un contrat *Automobile*.

Les appétences des individus pour des produits différents expliquent les différentes formes de distribution.

Application

Une application du modèle est réalisée sur une base de prospects en automobile. L'idée est d'identifier les clients à potentiels de multi-équipement et de leur proposer un rabais sur leur prime afin d'augmenter leur probabilité d'entrer en portefeuille. Certains de ces clients étant amenés à réaliser leur potentiel en se multi-équipant durant l'année. L'objectif de cette application est d'optimiser la valeur du portefeuille au bout d'un an (celle-ci correspond à la somme des valeurs des clients).

Les différents paramètres à prendre en compte dans cette application sont : le **rabais à accorder**, le **nombre de clients** bénéficiant de l'offre et l'évolution du **taux de transformation**.

- Le premier scénario réalisé est le **scénario témoin** dans lequel aucun prospect ne bénéficie de l'offre. Dans ce cas, le taux de transformation est de 50 %.
- Dans le second scénario, la moitié des clients possédant la probabilité de multi-équipement la plus forte se voit offrir un rabais sur leur prime équivalent à une baisse de valeur de 30 % pour leur premier contrat (55 euros en moyenne). Le taux de transformation de ces clients est de 75 %.
- Dans le scénario trois, le même rabais que précédemment est accordé cette fois-ci à l'intégralité de portefeuille et le taux de transformation reste inchangé.

Le tableau ci-dessous résume la valeur du portefeuille (en milliers d'euros) en moyenne au bout d'un an :

Scénario	Valeur totale du portefeuille	Valeur du 1 ^{er} contrat	Valeur dû au multi-équipement
Scénario Témoin	1 003	780	222
Scénario 2	1 051	755	295
Scénario 3	1 029	697	332

Ainsi, cette application montre qu'il est possible d'optimiser la valeur d'un portefeuille en se basant sur le critère du potentiel de multi-équipement.

Executive summary

Issue of this paper

Customer value is an indicator of customer profitability. It is composed of the **value of the contracts** and the **potential value** of the client.

- The value of the contracts corresponds to the projection of the technical and financial margins of each of the insurance products.
- The potential value is the value of future contracts weighted by the probability of subscription of these contracts.

Currently the potential value model is the same for all the company's customers. Thus there is no clear differentiation between individual and professional clients. The purpose of this thesis is **to develop a model solely around professional clients** in order to identify new segments of potential clients.

The model that will be developed is divided into two parts :

1. First of all, a probability of multi-equipment is calculated using statistical learning methods.
2. Then the distribution of the value in case of multi-equipment is calculated.

Modeling

Model the probability of multi-equipment

The data used in this paper to build the model correspond to the professionals in the portfolio from 2014 to 2016. The database of professionals in the portfolio in 2017 is used as a test sample for the model. This makes it possible to check the stability of the probabilities over time.

The probability of multi-equipment is calculated using two competing models : a logistic regression and a gradient boosting.

1. The logistic regression model :

The logistic regression model is sparse since few explanatory variables are used. This selection of variables is carried out via a Lasso penalty.

The selected model has 20 explanatory variables.

The variables that push the predicted probability most towards high values are :

- the fact of having at least one *Motor* contract ;
- the fact of having at least one *Motor Fleet* contract.

The variables that push the predicted probability most towards low values are :

- belonging to the distribution network '*Coutiers ou assimilés*' which corresponds to the network of brokers and similar entities (excluding Large Brokers) ;
- the fact of being single-equipped with *Life* or *Health* insurance product ;
- the fact of having a strong seniority.

2. The gradient boosting model :

This type of model is generally more efficient because it allows a better fit over the interactions between the different explanatory variables. However, it has the disadvantage of being difficult to interpret. Moreover, some parameters are not optimized by the algorithm, these are hyperparameters. Two of them, the depth and number of trees, are determined by cross-validation.

Feature importance statistic is used to determine the explanatory variables that provide the highest information gains to the model.

The variables that have the greatest impact are :

- (a) *the number of Automobile contracts* ;
- (b) *the seniority of the most recent contract* ;
- (c) *the client's seniority* ;
- (d) *the number of contracts held*.

3. Comparison of models :

The main performance statistic of the model that is used is the area under the ROC curve, it is summarized in the table below :

Model	2014 - 2016 Database	2017 Database
Logistic regression	74 %	79 %
Gradient boosting	78 %	81 %

Performance increases for the 2017 database because it has a lower proportion of multi-equipped individuals. However, the ROC curve is an indicator that is sensitive to unbalanced bases. However, the performance observed is better in both cases for the gradient boosting model.

We have also chosen to graphically represent the performance in order to judge the models in a qualitative way. On this graphical representation, the predictions are sorted and

grouped into packages of 3000. Within each grouping, the average of the predictions and the average of the true ones are calculated. In Figure 1, the green curve corresponds to the average of the true values and the black curve corresponds to the average of the predictions. Finally, the blue line corresponds to the average rate of multi-equipment over the entire database.

This graphical representation allows the volatility of predictions to be measured against actual values. Having 3000 individuals per group reduces volatility in the calculation of averages.

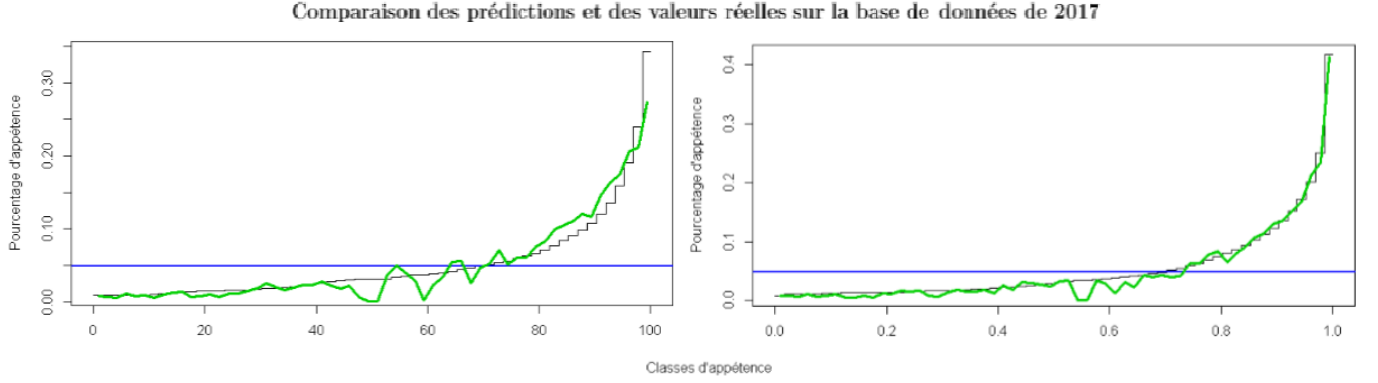


Figure 3 – **Left : logistic regression, right : gradient boosting**

For the rest of the model, it was decided to keep the gradient boosting model, which offers better performance.

Model profitability in the event of underwriting

The profitability in the event of subscription is modelled as the average value of the subscribed product multiplied by the probability of equipping said product :

$$\text{Average underwriting value} = \sum_{i \in \text{Products}} p_i \times \bar{V}_i$$

- **Construction of p_i :**

The probability of subscribing to product i , knowing that the customer is getting equipped, corresponds to the percentage of customers equipping themselves to this product. In addition, customers are divided into groups according to their nature (professional or company) and the products already owned.

The different products considered are :

Automobile except special machines, Home, special machines, merchant insurance, savings and pensions, Health and Life insurance and other PC products.

Thus we arrive at the different probabilities summarized in the table below :

Classe du client	Probabilité de souscrire le produit :						
	Auto hors E.G.	Autre IARD	Engins Spéciaux	Epargne Retraite	MRC	MRH + PNO	Prévoyance
Non PRO - IARD 00	24%	26%	4%	1%	23%	19%	4%
Non PRO - IARD 01	12%	57%	9%	1%	13%	8%	1%
Non PRO - IARD 10	53%	18%	9%	1%	14%	4%	2%
Non PRO - IARD 11	38%	27%	19%	1%	9%	3%	2%
Non PRO - Mixte	39%	23%	9%	4%	13%	6%	6%
Non PRO - Assurance de Pers.	12%	14%	0%	10%	14%	2%	47%
Non PRO - Assurance Vie	4%	5%	0%	73%	7%	0%	12%
PRO - IARD 00	43%	13%	3%	1%	8%	18%	15%
PRO - IARD 01	29%	25%	15%	0%	4%	17%	10%
PRO - IARD 10	52%	9%	5%	0%	3%	23%	7%
PRO - IARD 11	43%	13%	12%	0%	4%	22%	6%
PRO - Mixte	36%	10%	4%	3%	3%	23%	21%
PRO - Assurance de Pers.	13%	5%	1%	5%	2%	11%	63%
PRO - Assurance Vie	10%	1%	0%	27%	1%	2%	60%

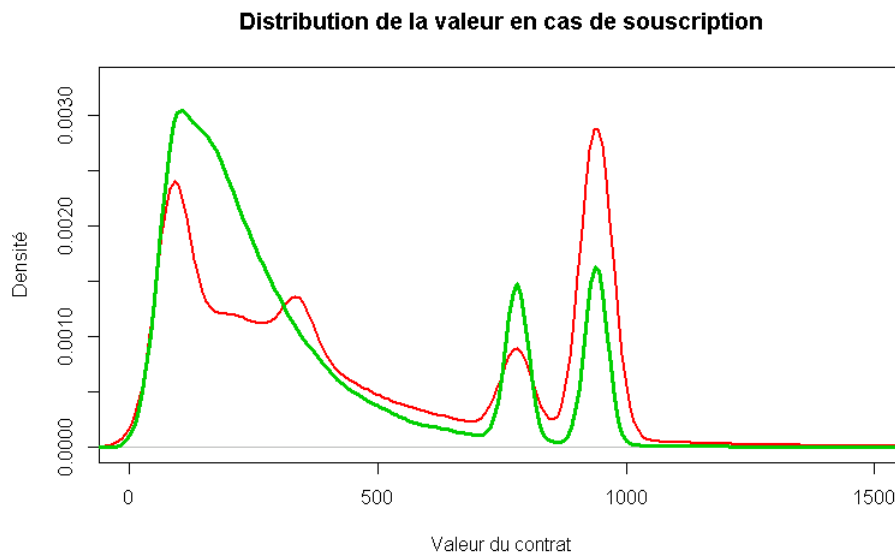
Figure 4 – Probability of underwriting equipped products

• Construction of $\bar{V}_i$:

- For products with a low underwriting volume : $\bar{V}_i$ is the average value of new business on product i .
- For products with a high subscription volume (*Automobile except special machines, Home and special machines*) : sub-populations are identified via regression trees in which the average value of contracts is calculated.

As the volatility remains high within the groups determined by the regression trees, it was decided to adjust a probability law to quantify this hazard.

Thus, thanks to the adjustment of these laws, we have the distribution of profitability in the event of underwriting as a law that combines previously adjusted laws and Dirac laws for products with low underwriting. The underwriting probabilities corresponding to the weights of each of the laws.



Above are represented the distributions of profitability in case of subscription for two different individuals :

- The red individual is a company with *Life* and *Non-Life* contracts.
- The green individual is a professional with a *Motor* contract.

Individuals' appetite for different products explains the different forms of distribution.

Application

An application of the model is carried out on a database of car prospects. The idea is to identify customers with multi-equipment potential and offer them a discount on their premium in order to increase their probability of entering the portfolio. Some of these customers are led to realize their potential by multi-equipping themselves during the year. The objective of this application is to optimize the value of the portfolio after one year (this corresponds to the sum of the clients' values).

The different parameters to be taken into account in this application are : the **discount to be granted**, the **number of customers** benefiting from the offer and the evolution of the **transformation rate**.

- The first scenario realized is the **control scenario** in which no prospect benefits from the offer. In this case, the transformation rate is 50 percent.
- In the second scenario, half of the customers with the highest probability of multi-equipment are offered a discount on their premium equivalent to a 30% decrease in value for their first contract (55 euros on average). The conversion rate for these customers is 75%.
- In scenario three, the same discount as before is granted this time to the entire portfolio and the conversion rate remains unchanged.

The table below summarises the value of the portfolio (in thousands of euros) on average after one year :

Scenario	Total value of the portfolio	Value of the first contract	Value due to multi-equipment
Control scenario	1 003	780	222
Scenario 2	1 051	755	295
Scenario 3	1 029	697	332

Thus, this application shows that it is possible to optimize the value of a portfolio based on the criterion of multi-equipment potential.

Remerciements

J'ai effectué mon apprentissage en alternance au sein de l'équipe « Valeur Client & Data Lab » de Generali. Je tiens à remercier toutes les personnes qui ont contribué à son bon déroulement et qui m'ont aidé lors de la rédaction de ce mémoire.

Je remercie tout d'abord Monsieur Simon Blaquière, directeur du département Données et Approche Client, ainsi que Monsieur Alexandre Cochard, manager de l'équipe « Valeur Client & Data Lab » pour leur accueil au sein de leurs équipes et leur encadrement bienveillant tout au long de mon travail à leur service.

Je souhaite aussi remercier Madame Louise Portas, manager de l'équipe Prévoyance collective, qui m'a accompagné durant la première partie de mon alternance, ainsi que Madame Samia Nouar, responsable d'études actuarielles, qui a repris le flambeau du tutorat. Je remercie également ma tutrice académique Madame Claire Boyer, maître de conférences au sein du Laboratoire de Probabilités, Statistique et Modélisation (LPSM) de Sorbonne Université. Toutes trois ont pu se rendre disponibles pour répondre à mes sollicitations tout au long de mon travail et je leur en suis très reconnaissant.

J'aimerais aussi remercier l'Institut de Statistique de l'Université de Paris (ISUP) pour la formation d'excellence reçue en actuariat, qui m'a permis d'étudier les notions utilisées dans ce mémoire.

Enfin, je remercie vivement l'ensemble de l'équipe « Valeur Client & Data Lab », avec qui j'ai eu le plaisir de travailler, pour leur disponibilité et le partage de leur expertise au quotidien.

Table des matières

Résumé	1
Abstract	2
Note de synthèse	3
Executive summary	8
Remerciements	13
Introduction	17
1 La valeur du client et le potentiel de valeur des clients professionnels	20
1.1 La modélisation de la valeur du client et son contexte	20
1.1.1 La valeur du client : une modélisation prospective de la rentabilité du client	20
1.1.2 La valeur potentielle	24
1.1.3 La valeur du client	25
1.2 Enjeu du mémoire	25
1.2.1 Le potentiel des clients professionnels	25
1.2.2 Nouvelle approche de la valeur potentielle	26
2 Construction des données et statistiques sur les données	28
2.1 Définition du périmètre	28
2.2 Analyse descriptive des données	28
2.2.1 Détermination du périmètre d'ancienneté des clients	28
2.2.2 Statistiques sur les produits d'appel	30
2.3 Construction de la base pour le modèle d'appétence au multi-équipement	36
2.3.1 Horizon de modélisation et fenêtre d'observation des données	36
2.3.2 Les variables explicatives	36
2.3.3 Modification des variables explicatives	37
2.3.4 Statistiques sur les variables explicatives	38
2.4 Construction de la base pour la rentabilité moyenne des contrats	48
2.4.1 Construction de la base d'apprentissage	48
2.4.2 Statistiques sur la variable réponse	49
2.4.3 Statistiques sur les variables explicatives	50
3 Modéliser la probabilité de souscription d'un nouveau contrat à horizon d'une année	54
3.1 La régression logistique	54
3.1.1 La famille exponentielle	54

3.1.2	Le modèle linéaire gaussien	55
3.1.3	Le modèle linéaire généralisé (GLM)	56
3.1.4	Estimation des paramètres	56
3.1.5	Pénalisation Lasso et Ridge	57
3.1.6	Les limites du modèle de régression logistique	59
3.2	Le gradient boosting	60
3.2.1	Pourquoi un tel modèle ?	60
3.2.2	Les arbres de classification et de régression	60
3.2.3	Le boosting et le gradient boosting	63
3.2.4	Les hyper-paramètres du modèle	64
3.2.5	L'importance des variables dans un modèle de gradient boosting	65
3.2.6	La valeur de Shapley et l'interprétation des boîtes noires	65
3.3	Comparaison des modèles	67
3.3.1	Indicateurs statistiques	67
3.3.2	Indicateurs de performance	69
3.3.3	Mesurer la pertinence des prédictions de façon graphique	71
3.4	Application et comparaison des modèles	71
3.4.1	Le modèle de régression logistique	72
3.4.2	Le modèle de gradient boosting	78
3.4.3	Conclusion	83
4	Modélisation de la rentabilité potentielle du contrat	85
4.1	De la modélisation de la valeur moyenne à la modélisation de la distribution . . .	85
4.2	Choix des produits dont la densité sera modélisée et probabilité d'équipement	86
4.2.1	Détermination des produits dont la densité de la valeur est modélisée	86
4.2.2	Détermination du panier produit en cas de souscription	89
4.3	Modélisation de la distribution de valeur par client	90
4.3.1	Création de classes de valeurs via des arbres de régression	90
4.3.2	Modélisation de la distribution de la valeur par classe	95
4.3.3	Illustration du modèle de rentabilité potentielle	101
4.3.4	Conclusion	103
5	La valeur potentielle et son application	104
5.1	Application à la politique de rabais	104
5.1.1	Idée de l'application	104
5.1.2	Hypothèse de l'application	105
5.2	La base de données utilisée	105
5.3	Résultats de l'application	107
5.3.1	<i>Scénario témoin</i>	107
5.3.2	Applications des scénarii accordant des rabais	107
5.4	Conclusion	112
	Conclusion	113
	Bibliographie	114
	Annexes	116

A	Intervalle de confiance sur les probabilités de souscription d'un produit	116
B	Graphiques d'ajustement des lois de probabilité pour chaque feuille des arbres de régression	118
B.1	Feuilles de l'arbre pour les produits <i>MRH & PNO</i>	118
B.2	Feuilles de l'arbre pour les produits <i>Automobile hors Engins Spéciaux</i>	131
B.3	Feuilles de l'arbre pour les produits <i>Engins Spéciaux</i>	145
C	Lien entre la prime et la valeur d'un contrat	150

Introduction

Cadre général : les évolutions récentes du marché assurantiel

La France est le premier marché européen de l'assurance, avec un chiffre d'affaires de 220 milliards d'euros en 2018. L'activité se répartit entre **l'assurance vie** (140 M€), **l'assurance maladie et les accidents corporels** (24 M€) et **l'assurance de biens et de responsabilité** (56 M€).

Les produits d'appel en assurance constituent aujourd'hui un marché très concurrentiel. Les compagnies ne peuvent plus appliquer d'importantes majorations à leurs clients sous peine d'accroître leur risque de résiliation. De plus, **la loi Hamon du 17 mars 2014 relative à la consommation**, permet à l'assuré de résilier son contrat à tout moment après un an de couverture et facilite donc la résiliation des contrats d'assurance. Ainsi, les assurés ont un moyen supplémentaire de faire jouer la concurrence entre compagnies.

Outre le fait que ce marché compétitif impose la compression des marges, les compagnies proposent des produits très similaires. En effet, certaines assurances sont à caractère obligatoire et présentent des garanties réglementaires. **Ces différentes raisons incitent les assureurs à travailler sur une vision transverse du client** et non pas seulement sur une vision par contrat ou par branche.

L'émergence de la notion de valeur du client

C'est dans ce cadre que la notion de valeur du client a été développée. Elle consiste à calculer la rentabilité projetée des contrats du client et son potentiel de valeur via la souscription de nouveaux contrats. **L'objectif est alors d'identifier les clients à forte valeur ou à fort potentiel, afin de les placer au cœur de la stratégie.**

Cette identification des clients à forte valeur ou à fort potentiel permet à la compagnie de mener des actions de fidélisation (par des avantages tarifaires par exemple), d'attaquer de nouveaux marchés et d'optimiser les offres sur les cibles de clients identifiées.

La valeur client chez Generali France

Generali est un groupe d'origine italienne indépendant et bénéficiant d'une forte présence à l'international. Fondé en 1831 et implanté en France dès l'année suivante où il figure aujourd'hui parmi les principaux assureurs. Generali est implanté dans plus de 50 pays et a réalisé un chiffre d'affaires de 66 milliards d'euros en 2018. Avec **61 millions de clients à travers le monde** et 71 000 collaborateurs. Notamment, le groupe Generali figure parmi les leaders sur les marchés d'Europe de l'ouest.

En France, le chiffre d'affaires de Generali a atteint 12,3 milliards d'euros en 2018. Il propose des solutions d'assurances et de gestion d'actifs à 7,2 millions de personnes ainsi qu'à **750 000 professionnels et entreprises**. Generali France s'appuie sur **plus de 9 000 collaborateurs et agents généraux**.

L'équipe **Valeur Client & Data Lab** appartient au département **Données et Approche client** et est composée d'un pôle « Data Lab » et d'un pôle « Valeur Client ». Le pôle « Valeur Client » compte **cinq collaborateurs**. Il s'agit d'une équipe encore récente en plein développement. Ses missions sont :

- le **calcul de la valeur du client** ;
- l'application de cet indicateur dans la **politique de souscription** et dans la **vie du contrat** ;
- la production de **reportings** sur la valeur du client à destination des autres services.

Objectif et problématique du mémoire

Les clients professionnels constituent une population différente à bien des égards des clients particuliers. En effet, leur volume de primes et le nombre de produits qu'ils détiennent sont plus élevés en moyenne que les clients particuliers. Il s'agit ainsi d'une population très profitable pour l'assureur.

Notre problématique est : **comment mesurer la rentabilité potentielle des clients professionnels ?**

Pour ce faire, un modèle de valeur potentielle spécifique aux clients professionnels sera développé. La valeur potentielle correspond à la valeur des contrats qu'un client pourrait être amené à souscrire dans le futur.

Cette problématique est pertinente d'un point de vue actuariel car elle permet de mettre en application, d'une part, des modèles statistiques de scoring afin d'identifier les clients qui ont le plus de chances de souscrire de futurs contrats. D'autre part, elle permet de mettre en application des méthodes d'ajustement de lois de probabilités afin d'identifier la distribution de la valeur des contrats.

De plus, une fois le potentiel des clients identifié, nous verrons comment il est possible d'optimiser la valeur des clients présentant du potentiel en agissant sur les offres commerciales.

Plan du mémoire

La première partie du mémoire explique le contexte autour de la valeur du client et détaille son calcul. De plus, elle explique l'approche qui va être utilisée pour le calcul de la valeur potentielle des clients professionnels (**chapitre un**).

Dans le **chapitre deux**, est réalisée une étude sur le potentiel des clients et sont présentés les bases de données qui seront utilisées par la suite.

Le **chapitre trois** traite de la modélisation de la probabilité de souscription d'un nouveau contrat et le **chapitre quatre** présente la modélisation de la valeur en cas de souscription.

Enfin, une application de la valeur potentielle dans l'attribution des offres commerciales est réalisée dans le **chapitre cinq**.

Chapitre 1

La valeur du client et le potentiel de valeur des clients professionnels

1.1 La modélisation de la valeur du client et son contexte

1.1.1 *La valeur du client : une modélisation prospective de la rentabilité du client*

La modélisation de la valeur du client chez Generali fait intervenir deux composantes :

- la valeur des contrats de l'assuré ;
- la rentabilité potentielle de l'assuré, via le potentiel de multi-équipement.

Dans la suite, nous détaillerons le calcul de cette notion encore récente dans les compagnies d'assurance.

1.1.1.1 ANALOGIE AVEC L'EMBEDDED VALUE

Le calcul de la valeur du contrat à Generali s'inspire de l'Embedded Value, qui est une mesure de la valeur des compagnies d'assurance vie du point de vue de l'actionnaire. Cet indicateur permet de répondre à une problématique spécifique à l'assurance vie. En effet, les contrats d'épargne-retraite ont vocation à rester en portefeuille durant plusieurs dizaines d'années. Ainsi, les investissements qui y sont liés se font sur le long terme et il est nécessaire de réaliser une modélisation prospective de la vie de ces contrats.

De fait, les résultats sur une année d'un portefeuille d'assurance vie ne sont pas représentatifs de la valeur réelle de celui-ci. Par exemple, les portefeuilles d'assurance vie sont déficitaires durant leur première année de commercialisation en raison des commissions et des frais de souscription (**figure 1.1**).

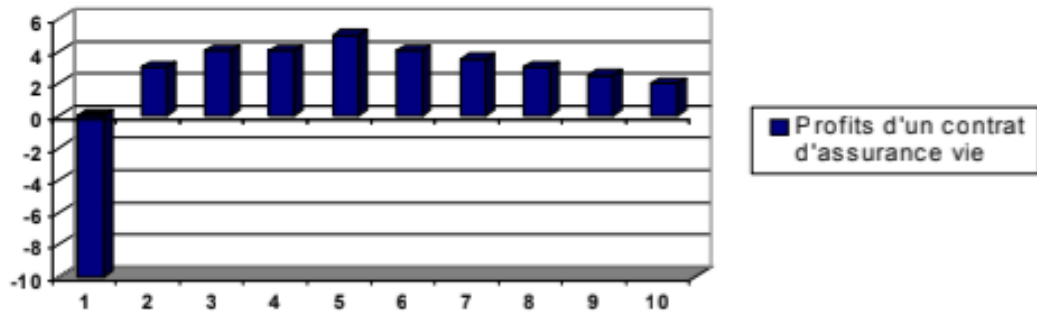


Figure 1.1 – Profits par année d'un contrat d'assurance vie

Source : Mathilde Garotin, *Market-consistent embedded value* (MCEV), ISUP, 2018-2019

Ainsi dans le contexte du marché sous tension des assurances et dans le but de fidéliser le client, il faut considérer que celui-ci, comme pour un portefeuille d'assurance vie, a vocation à rester plusieurs années client de la compagnie. **Il convient alors d'appliquer des pratiques analogues à l'Embedded Value pour déterminer sa valeur, en utilisant des hypothèses non plus uniquement à l'échelle du portefeuille, mais aussi à l'échelle du client.**

1.1.1.2 LA VALEUR DU CONTRAT

La **valeur du contrat** est la projection des marges techniques et financières estimées du contrat sur un horizon pouvant aller jusqu'à trente ans.

Le résultat technique s'écrit :

$$\text{Prime} - \text{Sinistre} - \text{Frais} - \text{Commissions}$$

Le résultat financier s'écrit :

$$\text{Provision} \times \text{Taux de rendement} - \text{Marge de solvabilité}$$

Les provisions constituent l'engagement de l'assureur vis-à-vis de l'assuré. La marge de solvabilité est le capital réglementaire que doit immobiliser l'assureur afin d'absorber les chocs subis par les provisions et de faire face aux aléas inhérents à l'activité d'assurance.

1. Le résultat technique

Les hypothèses et quantités utilisées pour la projection du résultat technique sont :

- le SP cible (i.e. : le ratio Sinistres / Primes) de la branche pour calculer la sinistralité de la première année ;
- l'inflation des sinistres (notée TX_{INFLA}) ;

- les majorations des primes (notées TX_{MAJO});
- les taux de commission et de frais de gestion (notés TX_{COMM});
- le taux de chute de l'individu (probabilité que le client résilie son contrat) (noté TX_{CHUTE});
- le taux d'imposition (noté TX_{IMPOT});
- les taux d'actualisation (notés $Déflateur$).

Ainsi, le résultat technique, noté **RT**, s'écrit :

$$\text{Soient } \forall i \begin{cases} P^i : \text{la prime de la } i\text{-ème année} \\ S^i : \text{le sinistre de la } i\text{-ème année} \\ F^i : \text{les frais de la } i\text{-ème année} \end{cases}$$

$$\text{Avec } \begin{cases} S^1 = P^1 \times SP \\ F^1 = P^1 \times TX_{COMM} \end{cases}$$

$$\text{Et } \forall i \geq 2 \begin{cases} P^i = P^{i-1} \times TX_{MAJO}^i \times (1 - TX_{CHUTE}^i) \\ S^i = S^{i-1} \times TX_{INFLA}^i \times (1 - TX_{CHUTE}^i) \\ F^i = P^i \times TX_{COMM} \end{cases}$$

$$\text{Alors : } \mathbf{RT} = \sum_{i=1}^{\text{Année max.}} (P^i - S^i - F^i) \times Déflateur_i \times (1 - TX_{IMPOT})$$

2. Le résultat financier

Les hypothèses et quantités utilisées pour la projection du résultat financier sont :

- les taux d'actualisation (notés $Déflateur$);
- les taux de rendement (notés TX_{Rend});
- le coût du capital;
- la marge de solvabilité (notée $TX_{MargeSolva}$);
- le taux d'imposition (noté TX_{IMPOT});
- les taux PSAP (Provision pour Sinistre À Payer) (notés TX_{PSAP}).

Ainsi, le résultat financier, noté **RF**, s'écrit :

$$\text{Soient } \forall i \begin{cases} Provision^i : \text{la provision de la } i\text{-ème année} \\ MS^i : \text{la marge de solvabilité de la } i\text{-ème année} \end{cases}$$

$$\text{Avec } \forall i \geq 1 \begin{cases} Provision^i = \sum_{k=1}^i TX_{PSAP}^k \times S^{i-k+1} \\ MS^i = P^i \times TX_{MargeSolva} \times Coût_capital \end{cases}$$

$$\text{Alors : } \mathbf{RF} = \sum_{i=1}^{\text{Année max}} (Provision^i \times TX_{Rend}^i - MS^i) \times Déflateur_i \times (1 - TX_{IMPOT})$$

3. La valeur du contrat

Enfin, la valeur du contrat correspond à la somme des résultats techniques et financiers projetés.

$$\text{Valeur du contrat} = RT + RF$$

Le cas pratique d'un contrat *Automobile* permettra de visualiser les flux financiers mis en jeu dans le calcul de la valeur d'un contrat : le **graphique 1.2** ci-dessous représente les flux de rendement pour un contrat *Automobile* en portefeuille d'ancienneté inférieure à une année. Le **rendement net**, par rapport au rendement brut, actualise les flux et intègre le taux d'imposition et la probabilité de chute (c'est-à-dire la probabilité que le client sorte du portefeuille).

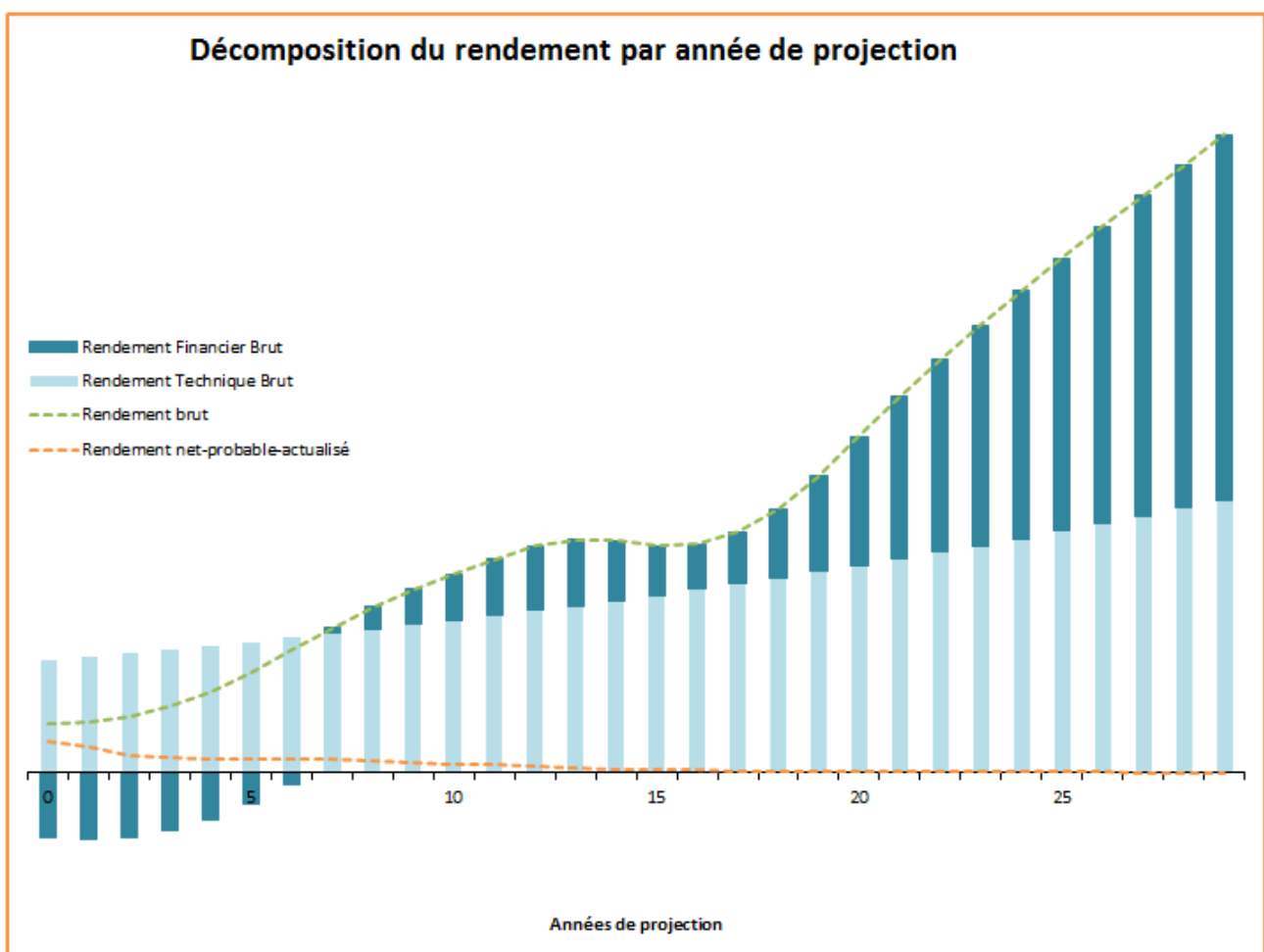


Figure 1.2 – Décomposition du rendement d'un contrat par année de projection

Plusieurs points sont à analyser concernant ce cas pratique de la valeur d'un contrat *Automobile*.

- Tout d'abord, le rendement technique brut évolue de manière linéaire croissante. En effet, le résultat technique suit les trajectoires de revalorisation prévues par les équipes du bureau d'étude technique.

- Concernant le rendement financier, il est négatif durant les premières années : cela s'explique par le fait que l'assureur doit immobiliser du capital. Ensuite, au fil des années, les provisions placées sur le contrat augmentent et cela a pour effet de faire croître le rendement financier.
- Enfin, le rendement net probable actualisé tend vers 0. Cela s'explique principalement par la loi de chute du client. À titre d'exemple, la probabilité qu'un client reste en portefeuille 10 ans est de l'ordre de 10 % — cette probabilité varie en fonction des caractéristiques du client (nature du client, produits détenus, etc.). En outre, le facteur d'actualisation accentue aussi la convergence vers 0, mais de façon plus marginale, puisque 1 euro dans 10 ans a une valeur actuelle inférieure (environ 90 centimes).

Finalement, la valeur du contrat correspond à la somme de chacun des rendements nets probables actualisés.

1.1.2 La valeur potentielle

À la valeur des contrats du client s'ajoute la **valeur potentielle du client**. Il s'agit de la rentabilité des nouveaux contrats que pourrait souscrire le client pondérée par la probabilité de souscription. Le modèle actuel calcule pour chaque client des scores de multi-équipement en fonction parfois du réseau de distribution. Generali possède trois principaux réseaux de distribution qui sont le réseau Agent, le réseau Courtier et le réseau Salarié.

Sept scores de multi-équipement ont été calculés via des méthodes de gradient boosting :

- probabilité de multi-équipement vers les contrats *Multirisques habitation (MRH)* pour un client détenteur de n'importe quel contrat (hors réseau salarié) ;
- probabilité de multi-équipement vers les contrats *MRH* pour un client détenteur de n'importe quel contrat (réseau salarié) ;
- probabilité de multi-équipement vers les contrats *Multirisques commerce (MRC)* pour un client détenteur d'au moins un contrat *Automobile* (hors réseau salarié) ;
- probabilité de multi-équipement vers les contrats *Epargne Retraite* pour un client détenteur d'au moins un contrat *Automobile* et/ou *MRH* (hors réseau salarié) ;
- probabilité de multi-équipement vers les contrats *Automobile* pour un client détenteur d'au moins un contrat *MRC* (hors réseau salarié) ;
- probabilité de multi-équipement vers les contrats *Automobile* pour un client détenteur d'au moins un contrat *MRH* (hors réseau salarié) ;
- probabilité de multi-équipement vers les contrats *Epargne Retraite* pour un client détenteur de n'importe quel contrat (réseau salarié).

La valeur potentielle du client est la somme, pour chacun des produits évoqués ci-dessus, des rentabilités moyennes par contrat pondérées par la probabilité de souscription :

$$\text{Valeur potentielle} = \sum_{i \in (\text{Auto}, \text{MRC}, \text{MRH}, \text{Eparg}/\text{Retr})} \text{Proba}_i * \bar{V}_i$$

Où $\bar{V}_i$ est la rentabilité moyenne du produit i .

La rentabilité $\bar{V}_i$ de ces contrats potentiels correspond à la moyenne de la valeur des contrats en portefeuille hors offre commerciale regroupés par catégorie (par exemple la tranche d'âge ou d'ancienneté de l'assuré, réseau, univers client, etc.).

Remarque :

Les scores ne sont calculés que pour des assurés n'ayant pas déjà le produit en question, ce qui peut amener à mal évaluer le potentiel dans la mesure où une part non négligeable du multi-équipement s'effectue sur le même produit (par exemple, plusieurs contrats *Automobile* ou plusieurs contrats *Multirisques habitation* pour un même assuré). Ce type de multi-équipement n'est pas scoré par le modèle actuel.

En résumé, il faut retenir que **la valeur potentielle est la rentabilité moyenne que l'on peut espérer d'un client sur les contrats qu'il peut souscrire, pondérée par sa probabilité de multi-équipement.**

1.1.3 La valeur du client

La **valeur du client** est le couple composé de la somme de la valeur de chacun des contrats de l'assuré et de sa valeur potentielle.

À partir de ce couple, des catégories de clients sont identifiées (clients à forte valeur, clients à potentiel, etc.).

1.2 Enjeu du mémoire

Dans la suite du mémoire, une valeur potentielle restreinte au périmètre des clients professionnels va être développée.

1.2.1 Le potentiel des clients professionnels

Les clients professionnels constituent une catégorie avec des besoins et des comportements différents des particuliers. Ils génèrent un chiffre d'affaire moyen beaucoup plus important, de l'ordre de 1 500 euros par an (contre 550 euros pour les particuliers).

De plus, certains clients professionnels sont fortement **multi-équipés** : un client est considéré comme multi-équipé s'il possède au moins deux contrats. Ce multi-équipement a pour effet notamment d'augmenter la durée du client et de ses contrats, et donc d'accroître la valeur du client. Ainsi, cette population est considérée comme ayant un potentiel important.

Dans l'approche actuelle de la valeur potentielle, les probabilités de multi-équipement et les valeurs moyennes des contrats sont calculées de la même façon sur tout le portefeuille. **L'objet du mémoire est de cerner au mieux le potentiel des clients professionnels en se focalisant uniquement sur ces derniers.**

1.2.2 Nouvelle approche de la valeur potentielle

Nous rappelons que la valeur potentielle est un indicateur mesurant l'appétence du client à se multi-équiper tout en quantifiant la rentabilité des potentiels contrats.

La valeur potentielle actuelle est l'agrégation de plusieurs valeurs potentielles par contrat (**figure 1.3**). Elle est donc difficile à appréhender telle quelle pour un service n'ayant pas connaissance du modèle et des probabilités sous-jacentes. Or la valeur potentielle, comme la valeur du client, a vocation à être utilisée par d'autres services.

Il a alors été choisi, dans cette modélisation de la valeur potentielle, de la séparer en deux composantes : la probabilité de multi-équipement et la distribution de la valeur en cas de souscription d'un nouveau contrat (**figure 1.4**). La probabilité et la distribution recouvrent tous les produits. Ainsi, le potentiel d'un client peut être appréhendé via trois grandeurs, qui sont :

- la probabilité de multi-équipement vers n'importe quel contrat ;
- la distribution de la rentabilité des contrats futurs de laquelle sont déduits des quantiles et la moyenne ;
- la valeur potentielle (qui correspond au produit des deux grandeurs précédentes).

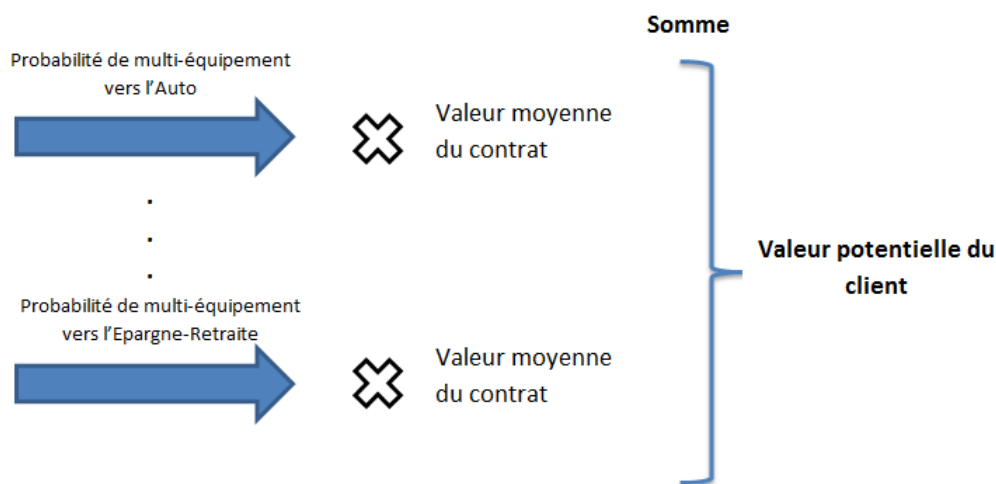


Figure 1.3 – Le modèle actuel de la valeur potentielle

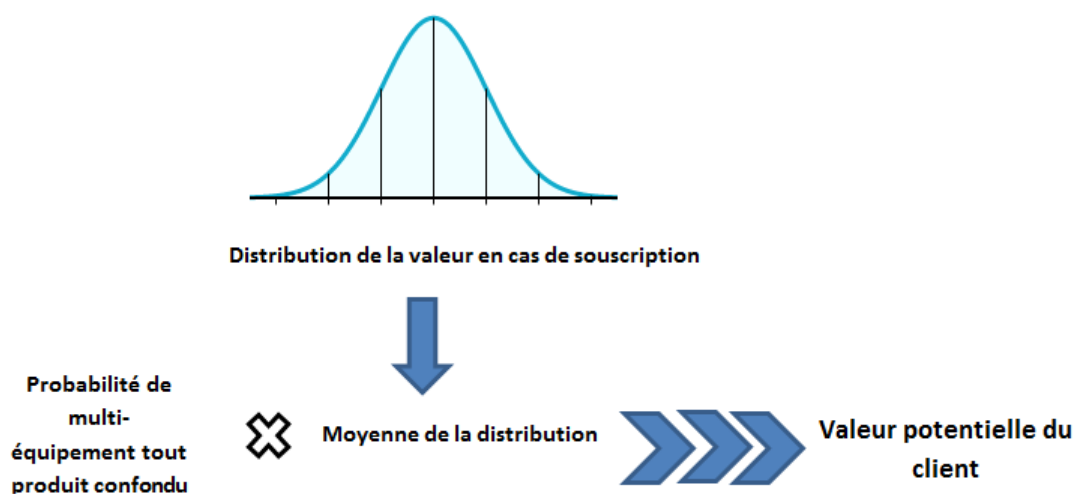


Figure 1.4 – Le modèle développé dans le mémoire

Concernant le calcul de la distribution de la rentabilité des contrats, des classes de valeurs seront identifiées via des arbres de régression puis la distribution au sein de ces classes sera estimée.

Pour rappel, dans la suite du mémoire, sera réalisée une étude statistique du potentiel de multiéquipement, puis seront présentées les bases de données utilisées pour les modèles ultérieurs (**chapitre 2**). Ensuite, seront présentés les modèles statistiques permettant la modélisation des différentes composantes de la valeur potentielle (**chapitres 3 et 4**). Enfin, une application de la valeur potentielle sur l'optimisation des offres commerciales sera proposée (**chapitre 5**).

Chapitre 2

Construction des données et statistiques sur les données

Le but de ce chapitre est d'introduire le concept de potentiel de multi-équipement et de présenter les bases de données qui vont être utilisées dans les modèles développés par la suite (**chapitres 3 et 4**).

Dans la première partie sera défini le périmètre de l'étude (**partie 2.1**). Ensuite, une analyse descriptive des données est réalisée, son but étant d'étudier les tendances de multi-équipement sur le portefeuille en fonction du produit d'appel (**partie 2.2**). Enfin, la base de données utilisée pour le modèle prédisant la probabilité de multi-équipement est présentée dans la **partie 2.3** et la base de données intervenant dans la modélisation de la valeur en cas de souscription est présentée dans la partie **partie 2.4**.

2.1 Définition du périmètre

Afin de traiter la problématique du calcul de la rentabilité potentielle des clients professionnels, il est nécessaire de restreindre l'étude aux clients dont la nature est :

- Association
- Copropriété
- Professionnel
- Petite entreprise (effectif inférieur à 20 salariés)
- Entreprise (effectif supérieur à 20 salariés).

De plus, nous travaillerons uniquement sur les réseaux Agents et Courtiers.

2.2 Analyse descriptive des données

2.2.1 Détermination du périmètre d'ancienneté des clients

Une première étude est réalisée sur la base de données regroupant les clients entrés en portefeuille entre 2010 et 2018. Cette base est construite en récupérant l'équipement du client à

la fin de chaque année. Ainsi, elle nous livre l'évolution du panier produit de chacun des assurés (**le panier produit** est l'ensemble des produits détenus par un client).

Cette base de données contient 715 548 assurés, dont **198 886 (28 %)** qui se multi-équipent. La représentation de l'entrée en multi-équipement, caractérisée par la souscription d'un second contrat, est donnée par la **figure 2.1**.

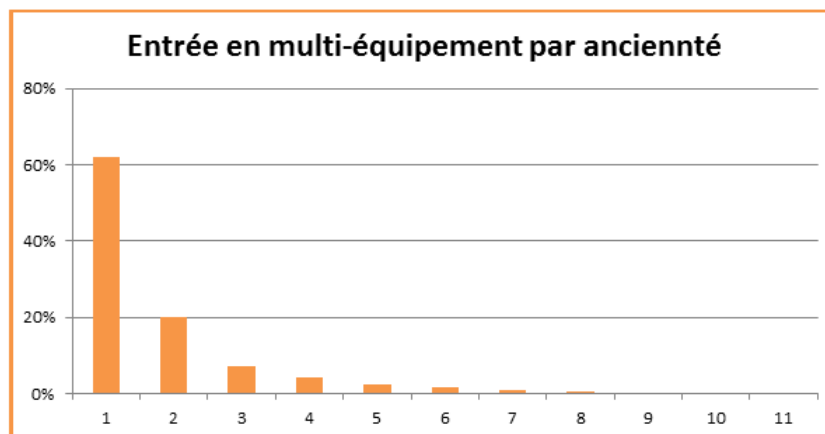


Figure 2.1 – Répartition de la souscription d'un second contrat en fonction de l'ancienneté

Lecture :

- Plus de 80 % de l'entrée en multi-équipement se fait durant les deux premières années dans le portefeuille.
- Passée cette période, les clients mono-équipés ont de faibles chances de se multi-équiper.

Remarque : le **graphique 2.1** ci-dessus ne représente toutefois que la souscription du second contrat, nous n'avons pas l'information concernant les contrats qui sont souscrits par les clients déjà multi-équipés. La question est alors de savoir si les clients déjà multi-équipés continuent à s'équiper après deux ans d'ancienneté.

Le **graphique 2.2** suivant présente une approche plus générale sur l'évolution de l'équipement des clients. Il comptabilise l'intégralité du multi-équipement réalisé par les clients (en vert) et pas seulement l'équipement d'un second contrat. Les clients qui s'équipent et se déséquipent sont représentés en volume et en pourcentage.

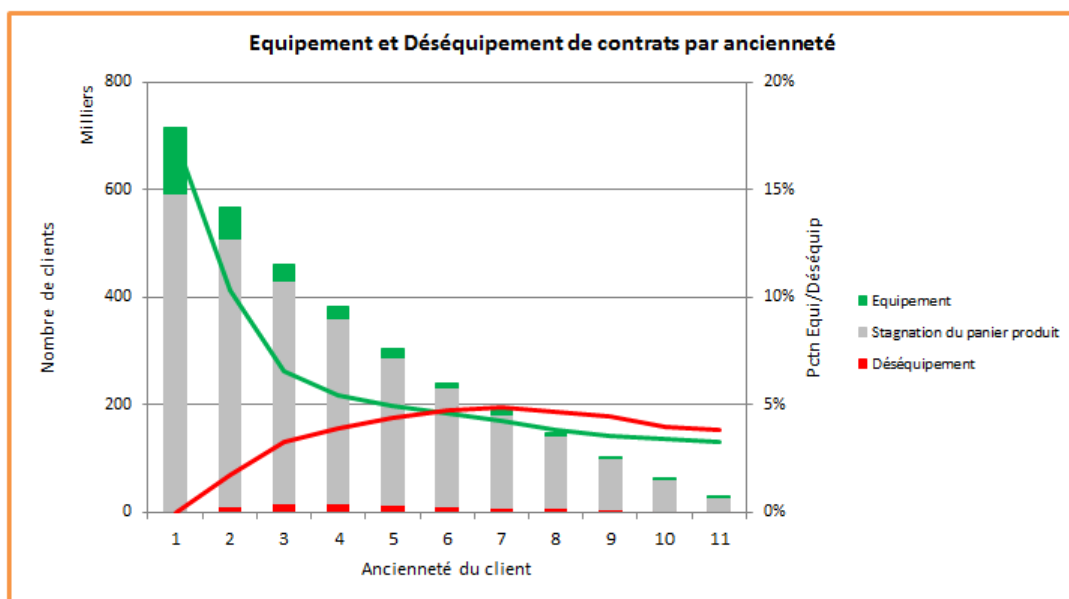


Figure 2.2 – Evolution de l'équipement des clients par ancienneté

Remarque : le déséquipement (en rouge) représente les clients en portefeuille dont le nombre de contrat diminue au cours de l'année mais qui restent en portefeuille. Les clients qui sortent complètement du portefeuille ont pour conséquence de réduire l'effectif du nombre de clients.

Lecture :

- Le pourcentage de multi-équipement est de 19 % la première année et chute à 6 % la troisième année.

Le taux de multi-équipement chute rapidement une fois passées les deux premières années d'ancienneté. **Ainsi le choix est fait, dans la suite, de ne modéliser que le potentiel de multi-équipement des clients de moins de deux ans d'ancienneté.** Au-delà, le potentiel des clients n'est plus modélisé.

2.2.2 Statistiques sur les produits d'appel

Une première analyse macroscopique du portefeuille est réalisée afin de mieux cerner les notions de multi-équipement et de potentiel de multi-équipement. Cette analyse est réalisée avec la même base que l'étude du paragraphe précédent et est articulée autour de la variable "*Produit d'appel*". Un **produit d'appel** est un produit par lequel le client va entrer en portefeuille : il s'agit du premier contrat que va souscrire le client.

Nous essaierons d'identifier des tendances de multi-équipement vis-à-vis de cette variable explicative.

2.2.2.1 LA RÉPARTITION DES PRODUITS D'APPEL

Le **graphique 2.3** ci-dessous représente la répartition des produits d'appel.

Des catégories de produits sont créées pour les produits ne possédant pas un poids significatif afin de rendre le graphique plus lisible. Par exemple, la catégorie *Epargne Retraite* regroupe les produits : *Retraite Individuelle*, *Retraite Collective* et *Epargne*.

Le demi-cercle extérieur détaille la répartition des produits présents dans un tel regroupement.

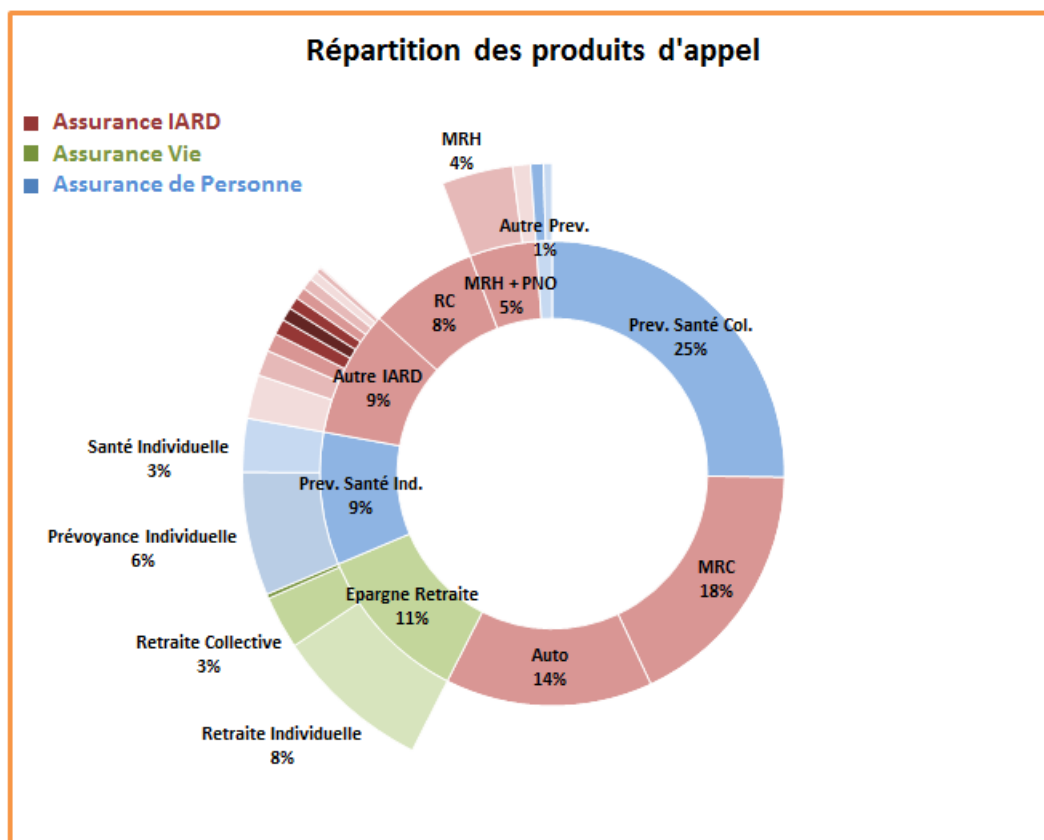


Figure 2.3 – Répartition du produit d'appel parmi les clients professionnels

Lecture :

- Les produits d'assurance **Incendies, accidents, risques divers (IARD)** sont majoritaires (*Multirisques commerce (MRC)*, *Automobile*, *Responsabilité civile (RC)*, *Multirisques habitation Propriétaire non occupant (MRH & PNO)*, autres : **53 %**), suivis de l'**assurance de personnes** (*Prévoyance santé collective, individuelle, autre* : **36 %**).
- Trois produits IARD réalisent 40 % de l'entrée en portefeuille (*Automobile*, *Responsabilité civile (RC)* et *Multirisques commerce (MRC)*), et à eux trois 75 % de l'entrée en portefeuille de produits IARD.

Ainsi, l'entrée dans le portefeuille est fortement dominée par les produits IARD, suivis ensuite des produits de Prévoyance santé collective. **Afin d'évaluer le potentiel de valeur lié au multi-équipement, il est nécessaire de savoir si d'une part il existe une corrélation entre les produits d'appel et le multi-équipement, et d'autre part de connaître la composition des paniers produits une fois passée la première année dans le portefeuille.**

Pour information, les regroupements de produits réalisés ainsi que les pourcentages correspondants pour "*Autre Prévoyance*" et "*Autre IARD*" sont les suivants :

- "Autre Prévoyance" :
 1. *Emprunteur* : 0,7 %
 2. *Garantie Accidents de la vie (GAV)* : 0,5 %
- "Autre IARD" :
 1. *Multirisques immeubles (MRI)* : 2,3 %
 2. *Construction* : 1,3 %
 3. *Garage et concessions* : 1 %
 4. *Transport* : 0,8 %
 5. *Risques industriels* : 0,7 %
 6. *Flottes automobiles* : 0,6 %
 7. *Multirisques agricoles (MRA) et Grêle* : 0,6 %
 8. *Risques techniques* : 0,6 %
 9. *Plaisance* : 0,5 %
 10. *Protection juridique* : 0,3 %
 11. *Risques divers* : 0,1 %

Enfin, dans la catégorie *Epargne Retraite*, le produit *Epargne* représente 0,2 % de l'entrée en portefeuille. Dans la catégorie *Multirisques habitation & Propriétaire non occupant (MRH & PNO)*, le produit *PNO* représente 1 % de l'entrée en portefeuille.

2.2.2.2 LE MULTI-ÉQUIPEMENT AU COURS DE LA PREMIÈRE ANNÉE

L'équipement des clients est maintenant observé après un an d'ancienneté. On considère qu'un client s'est multi-équipé au cours de sa première année en portefeuille si, à la fin de l'année, son nombre de contrats est supérieur ou égal à deux.

Cependant, un client qui souscrit plusieurs contrats lors de son entrée en portefeuille puis qui n'en souscrit plus a été jusqu'à présent inclus dans l'analyse. Or, il ne doit pas être considéré comme ayant du potentiel, puisqu'il ne souscrit plus aucun nouveau contrat. Dans la suite, lorsque nous feront référence à des clients qui se sont multi-équipés, seuls les clients qui ont réellement réalisé un potentiel seront considérés (i.e. les clients qui ne se multi-équipent que lors de l'entrée en portefeuille ne sont plus comptabilisés).

Le **tableau 2.4** ci-dessous représente le pourcentage de multi-équipement sur tout le portefeuille et par produit d'appel.

Produit d'entrée	% de clients multi-équipés
Tout produit confondu	14%
Auto	28%
Autre IARD	15%
Autre Prev.	24%
Epargne Retraite	8%
MRC	12%
MRH + PNO	28%
Prev. Santé Col.	7%
Prev. Santé Ind.	15%
RC	14%

Figure 2.4 – Taux de multi-équipement à la fin de la première année en portefeuille

Lecture :

- Les produits *Automobile*, *Autre Prévoyance* et *MRH & PNO* ont un taux de multi-équipement bien supérieur à la moyenne (**Tout produit confondu**).
- Les produits *Epargne Retraite* et *Prévoyance santé collective* ont des taux de multi-équipement près de deux fois inférieurs à la moyenne du portefeuille.

Remarque : La moyenne de multi-équipement est de 14 % contre 19 % sur le **graphique 2.2**. Cela s'explique par le fait que les clients qui se multi-équipent seulement lors de l'entrée en portefeuille ne sont plus comptabilisés comme étant des clients qui réalisent un potentiel de multi-équipement.

De plus, en combinant le taux de multi-équipement avec la répartition par produit d'appel (**graphique 2.3**), il est possible de déterminer la répartition des clients qui se multi-équipent en fonction du produit d'appel.

Par exemple, la part des clients multi-équipés entrant en portefeuille par le produit *Multirisques habitation Propriétaire non occupant (MRH & PNO)* est de :

$$\frac{\text{Taux de multi-équipement} \times \text{Pourcentage de clients entrant par le produit}}{\text{Taux de multi-équipement en moyenne}} = \frac{28\% \times 5\%}{14\%} = 10\%$$

Ainsi, 10 % des clients qui se multi-équipent entrent en portefeuille par le produit *MRH & PNO*. Le **tableau 2.5** ci-dessous présente les taux pour les autres produits.

Produit d'entrée	Répartition des clients multi-équipés en fonction du produit d'appel
Auto	28%
Autre IARD	9%
Autre Prev.	2%
Epargne Retraite	7%
MRC	16%
MRH + PNO	10%
Prev. Santé Col.	12%
Prev. Santé Ind.	9%
RC	8%

Figure 2.5 – Répartition du multi-équipement en fonction du produit d'appel

Lecture :

- 56 % des clients qui se multi-équipent sont entrés en portefeuille par des contrats *Auto-mobile*, *Multirisques commerce (MRC)* ou *Prévoyance santé collective*.

Ces taux sont intéressants car ils permettent de mettre en évidence **un fort effet volume : les trois produits d'appels qui génèrent le plus de multi-équipement sont aussi les trois produits d'appels qui font entrer le plus de clients**. Ainsi, un produit avec un taux de multi-équipement faible ou moyen peut malgré tout peser de façon conséquente dans la répartition totale des clients multi-équipés : c'est le cas des produits de *Prévoyance santé collective* ou des produits *Multirisques commerce (MRC)*.

Observons maintenant la distribution des produits qui sont multi-équipés. Le **graphique 2.6** ci-dessous représente la distribution des produits multi-équipés en fonction du produit d'entrée. Au sein de chaque bâton, chaque couleur représente le poids respectif de la famille du produit d'entrée (*Personnes*, *IARD*, *Vie*).

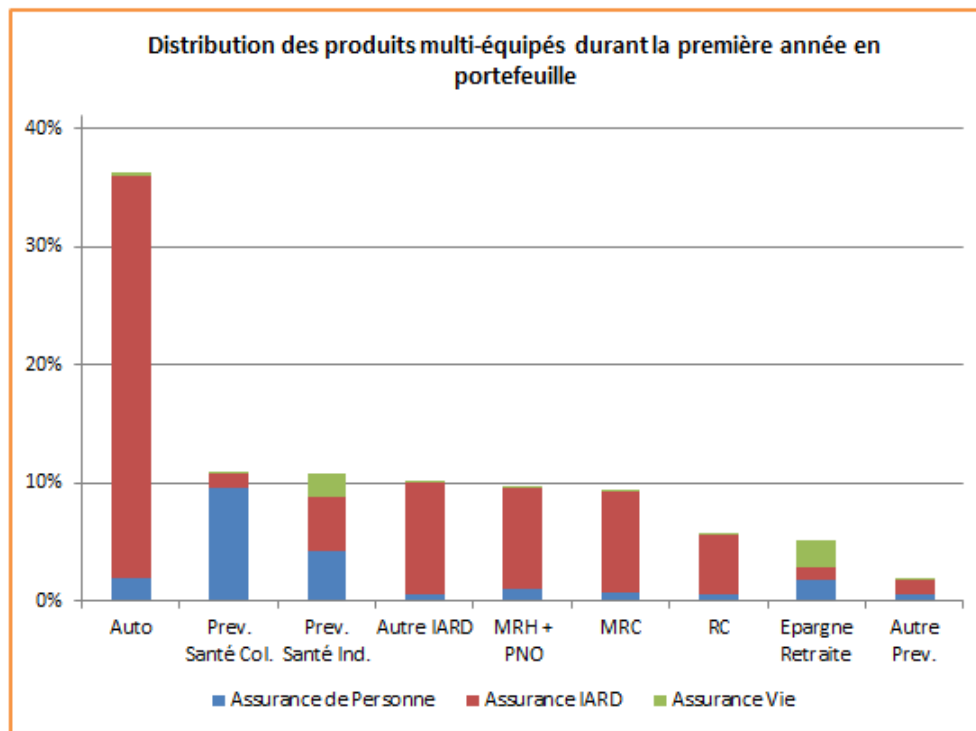


Figure 2.6 – Répartition des produits qui sont multi-équipés en fonction du produit d'entrée

Lecture :

- 36 % du multi-équipement se fait vers des produits *Automobile*.
- Les clients se multi-équipant en *Prévoyance santé collective* sont presque tous entrés dans le portefeuille via un premier contrat d'assurance de *Personnes*. Pour information, il s'agit à 95 % de clients entrés via la *Prévoyance santé collective*.

La majorité des clients qui se multi-équipent s'équipe dans la même famille de produits (*Personnes*, *IARD*, *Vie*) que dans celle par laquelle ils sont entrés en portefeuille.

À ce stade, il est possible de conclure que :

- Les clients ayant le plus fort potentiel de multi-équipement entrent par les produits *Automobile* et *MRH* et de façon plus générale par les produits *IARD*.
- La *Prévoyance santé collective* étant le premier produit d'appel (25 % des clients entrent par ce produit), ce produit bénéficie d'un effet volume qui ne doit toutefois pas être négligé.
- Le reste des produits d'assurance *Vie* et de *Personnes* est considéré comme présentant peu de potentiel (les clients entrés par la catégorie *Autre prévoyance* sont trop minoritaires pour que l'on puisse les considérer comme une population à potentiel).

2.3 Construction de la base pour le modèle d'appétence au multi-équipement

2.3.1 Horizon de modélisation et fenêtre d'observation des données

L'objectif est de déterminer la probabilité qu'a la client de souscrire un nouveau contrat, c'est-à-dire la probabilité qu'il possède un contrat supplémentaire dans son panier produit.

Nous fixons l'horizon de la probabilité à un an. Ce choix découle de raisons pratiques et stratégiques.

- D'une part, un horizon court d'un an permet d'augmenter la taille de la base de données et donc la pertinence statistique. En effet, si nous observons par exemple des données de 2014 à 2016, nous disposons de trois périodes d'observation pour un horizon d'un an (*Début 2014 - Fin 2014*, ..., *Début 2016 - Fin 2016*). Dans le cas d'un horizon de deux ans, nous ne disposerions plus que de deux périodes (*Début 2014 - Fin 2015*, *Début 2015 - Fin 2016*). Lorsque les statistiques sont réalisées sur trois ou quatre années, le gain d'une période d'observation peut être très significatif.
- D'autre part, nous essaierons par la suite de déterminer la valeur moyenne découlant du multi-équipement. Or, la valeur dépend beaucoup des tarifs en vigueur, donc l'horizon à un an apparaît être un bon compromis.

La fenêtre d'observation pour les données d'entraînement du modèle couvre 2015 à 2017. Une fois le modèle entraîné et validé (via une validation croisée), **celui-ci est testé sur les données de 2017-2018 afin de mesurer sa stabilité dans le temps.**

La base de données d'entraînement contient 387 077 individus et la base de données de 2017-2018 contient 179 475 individus.

2.3.2 Les variables explicatives

Les variables explicatives utilisées pour construire la base de données sont :

1. Les données signalétiques client :
 - le réseau de distribution ;
 - l'âge ;
 - l'ancienneté ;
 - la catégorie socio-professionnelle (CSP) ;
 - l'e-mail (indicatrice renseignant si la variable est complétée) ;
 - le nombre d'enfant(s) ;
 - le statut de l'occupant (locataire, propriétaire) ;
 - la banque ;
2. Les données de la situation courante à Generali :
 - la nature du client ;

- le nombre de contrats *IARD*, *Vie* et total ;
- le montant de prime nette (pour les catégories : *Assurance de personnes*, *Dommages aux biens*, *Automobile*, *MRC*, *MRH*, *Epargne Retraite*, *GAV*, *IARD*, *Vie*, total) ;
- le panier produit du client ;
- le nombre de sinistre(s) à 12 mois ;
- le produit d’appel du client ;
- le délai entre la souscription du premier et du second contrat ;
- le délai entre la souscription de l’avant dernier et du dernier contrat ;

3. Les données du contrat le plus récent du client :

- la réduction commerciale ;
- l’ancienneté du contrat ;
- le code territoire.

Enfin, la variable à expliquer est l’indicateur du multi-équipement au cours de l’année. S’il y a eu multi-équipement durant l’année, la variable vaut 1, sinon elle vaut 0. Cette variable est créée en comparant le nombre de contrats détenus par le client en N et $N+1$.

Remarque importante : le taux de multi-équipement dans la base de données d’entraînement est de **6,1 %**, ce taux est plus faible que celui vu dans la partie 2.2. Cela s’explique par le fait que la base est composée de tous les clients de moins de deux ans d’ancienneté en portefeuille. Ainsi, une part importante des clients s’est déjà multi-équipée (17 %) et a donc déjà réalisée son potentiel.

Le taux de multi-équipement de la base de données de 2017-2018 est de **4,9 %**.

2.3.3 Modification des variables explicatives

Une première transformation des variables est réalisée pour les rendre compatibles avec les modèles d’apprentissage statistique qui vont être utilisés. Les valeurs manquantes et le découpage en classes de certaines variables seront traités dans la suite.

Une première analyse révèle un fort de taux de valeurs non renseignées concernant les données signalétiques (**tableau 2.7**). Le choix est fait de transformer ces variables en indicatrices valant 1 si la variable est renseignée et 0 sinon. Ces variables mises sous forme d’indicatrices restent potentiellement utiles car elles peuvent traduire l’état de la relation entre l’intermédiaire et le client ou d’autres phénomènes.

Variable	% de valeurs non-manquantes
Email	39%
Banque	33%
Age	18%
Statut de l’occupant	13%
CSP	12%
Nombre Enfant	10%

Figure 2.7 – Pourcentage de variables correctement renseignées

Enfin, certaines variables numériques sont transformées en classes. Les variables sont découpées en deux classes selon la valeur de la médiane. Le **tableau 2.8** indique les variables qui ont été découpées.

Variable	Modalités	Pourcentage
Nombre d'enfant	Inférieur ou égal à 0	95%
	Supérieur à 0	5%
Prime nette totale	Inférieur ou égal à 235	49%
	Supérieur à 235	51%
Prime nette IARD	Inférieur ou égal à 0	50%
	Supérieur à 0	50%
Prime nette Vie	Inférieur ou égal à 0	87%
	Supérieur à 0	13%
Réduction commerciale	Inférieur ou égal à 0	80%
	Supérieur à 0	20%

Figure 2.8 – Variables découpées en classes

2.3.4 Statistiques sur les variables explicatives

Dans cette partie, le lien entre certaines variables explicatives et la variable réponse est analysé.

2.3.4.1 ANALYSE DE LA VARIABLE : MONTANT DE PRIME NETTE

La variable Montant de prime nette est découpée en classes de taille homogène. Chaque classe contient environ 20 000 individus, sauf la classe des individus ayant une prime nulle qui contient 100 000 individus. Cette classe est composée des contrats Epargne et Prévoyance Santé Collective. Dans le cas des contrats Epargne, aucune prime n'est renseignée puisqu'il s'agit de montants d'encours. Dans le cas des contrats Prévoyance Santé Collective cela est dû au fait que les primes ne redescendent pas dans les bases de données.

Le **graphique 2.9** représente le taux de multi-équipement au sein de chaque classe.

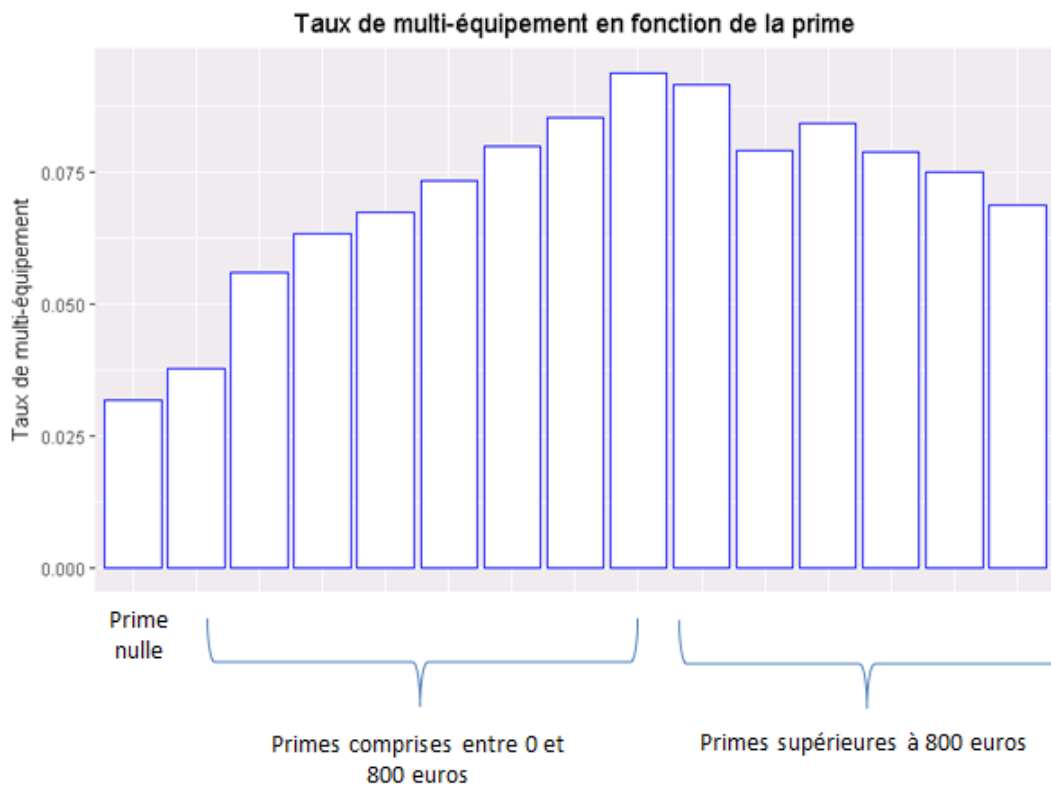


Figure 2.9 – Variables découpées en classes

Le taux de multi-équipement présente un comportement non linéaire. Il croît jusqu'aux primes d'un montant de 800 euros puis décroît. Ce type de comportement est difficilement modélisable dans le cas d'un modèle linéaire.

2.3.4.2 ANALYSE DES VARIABLES QUALITATIVES

Dans la suite de cette partie, nous traiterons seulement les variables qualitatives. Le nombre d'observations et le taux moyen de clients multi-équipés par modalité des variables explicatives est représenté.

La couleur bleue dans les graphiques ci-dessous représente les clients qui se sont multi-équipés alors que la couleur rouge fait référence aux clients qui ne se sont pas multi-équipés. Les losanges indiquent le taux de clients se multi-équipant, la lecture de ce taux se fait sur le second axe des ordonnées à droite.

- La **figure 2.10** représente la variable *Code Territoire* :

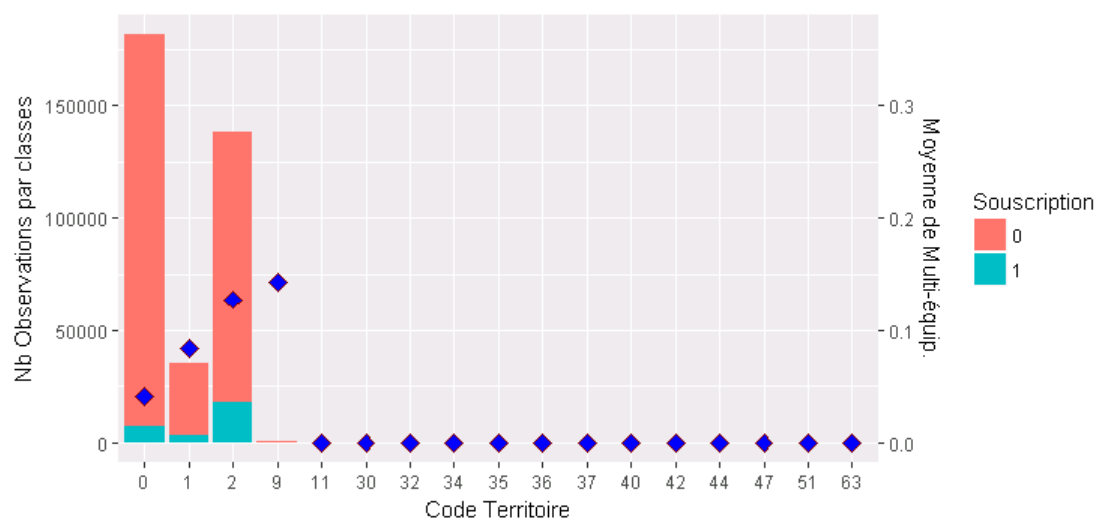


Figure 2.10 – Taux moyen de clients qui se multi-équipent en fonction de la variable *Code Territoire*

Détail des modalités de la variable :

- 0 : le *Code Territoire* est manquant ;
- 1 : région parisienne ;
- 2 : province ;
- 9 à 63 : étranger.

Du fait de la faible volumétrie des clients possédant un *Code Territoire* faisant référence à l'étranger, ces clients sont retirés de l'étude.

- La **figure 2.11** représente la variable *Type du réseau de distribution* :

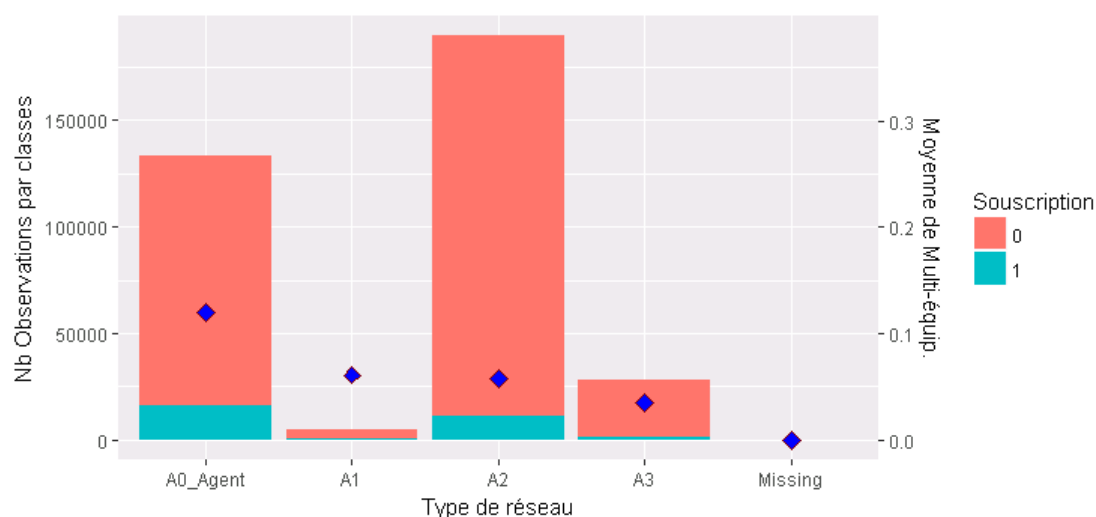


Figure 2.11 – Taux moyen de clients qui se multi-équipent en fonction de la variable *Type du réseau de distribution*

Détail des modalités de la variable :

- A0_Agent : réseau de distribution Agent ;
- A1 : réseau de distribution Grands Courtiers ;
- A2 : réseau de distribution Courtiers et assimilés ;
- A3 : autres réseaux ;
- Missing : la variable n'est pas renseignée.

Il est décidé de créer une nouvelle variable qui regroupe les modalités A1, A2 et A3. En effet, ces modalités présentent un taux proche de clients se multi-équipant. Ainsi, cette nouvelle variable oppose le réseau *Agent* aux autres réseaux.

Les deux variables sont laissées dans la base de données dans la mesure où des méthodes de sélections de variables seront effectuées.

- La **figure 2.12** représente la variable *Indicatrice du Statut de l'occupant* :

Cette variable vaut 1 si le statut de l'occupant (locataire, propriétaire, etc.) est renseigné, sinon elle vaut 0.

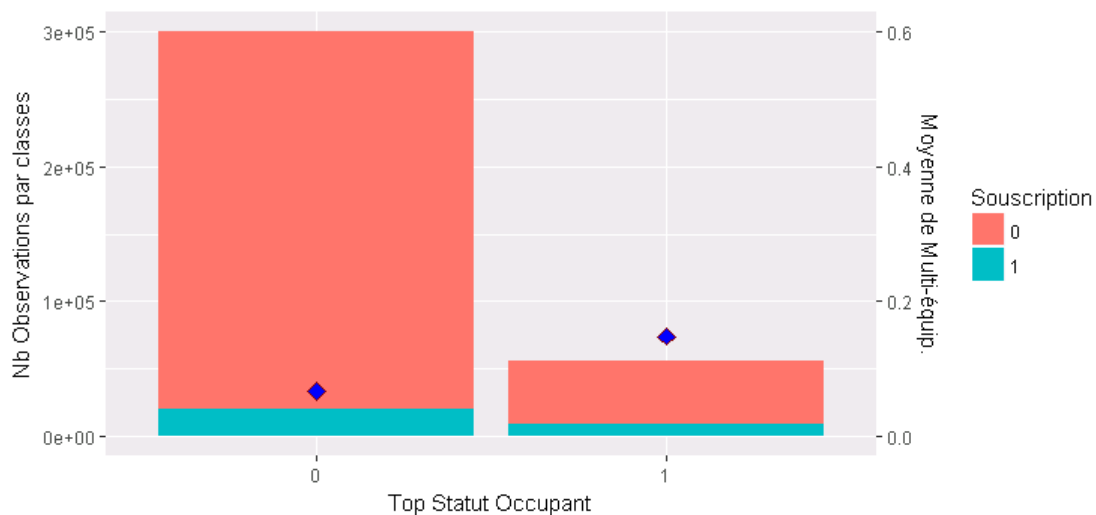


Figure 2.12 – Taux moyen de clients qui se multi-équipent en fonction de la variable *Indicatrice du Statut de l'occupant*

- La **figure 2.13** représente la variable *Indicatrice du Multi-équipement* :

Cette variable vaut 1 si le client est déjà multi-équipé (i.e. : le client possède plusieurs contrats), sinon elle vaut 0.

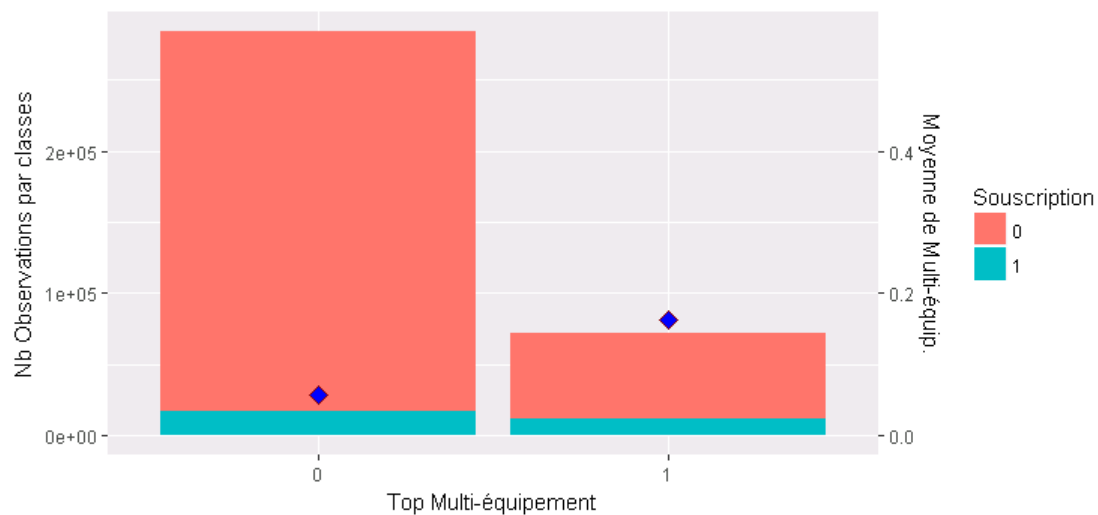


Figure 2.13 – Taux moyen de clients qui se multi-équipent en fonction de la variable *Indicatrice du Multi-équipement*

Ainsi, si un client possède plus de deux contrats, les chances qu'il se multi-équipe durant l'année sont de plus de 15 %.

- La **figure 2.14** représente la variable *Indicatrice de la possession d'au moins un contrat Automobile* :

Cette variable vaut 1 si le client possède au moins un contrat *Automobile*, sinon elle vaut 0.

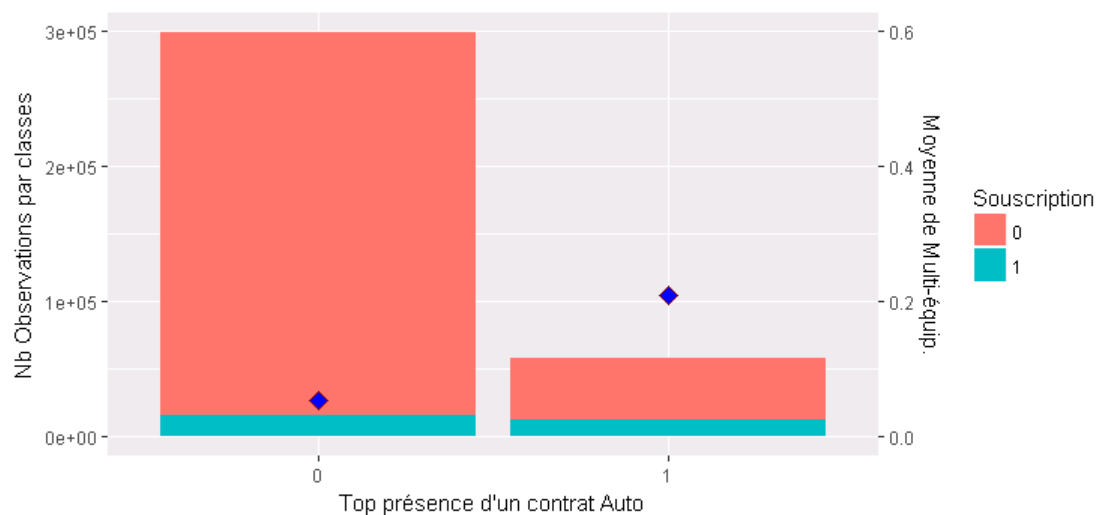


Figure 2.14 – Taux moyen de clients qui se multi-équipent en fonction de la variable *Indicatrice de la possession d'au moins un contrat Automobile*

Le taux des clients qui possèdent au moins un contrat *Automobile* et qui se multi-équipent est de plus de 20 %. Cela rejoint les conclusions de la **partie 2.2**, cependant le taux est ici plus faible car une partie des clients s'est déjà multi-équipée et a donc déjà réalisé son potentiel.

- La **figure 2.15** représente la variable *Indicatrice de la CSP* :
Cette variable vaut 1 si la CSP du client est renseignée, sinon elle vaut 0.

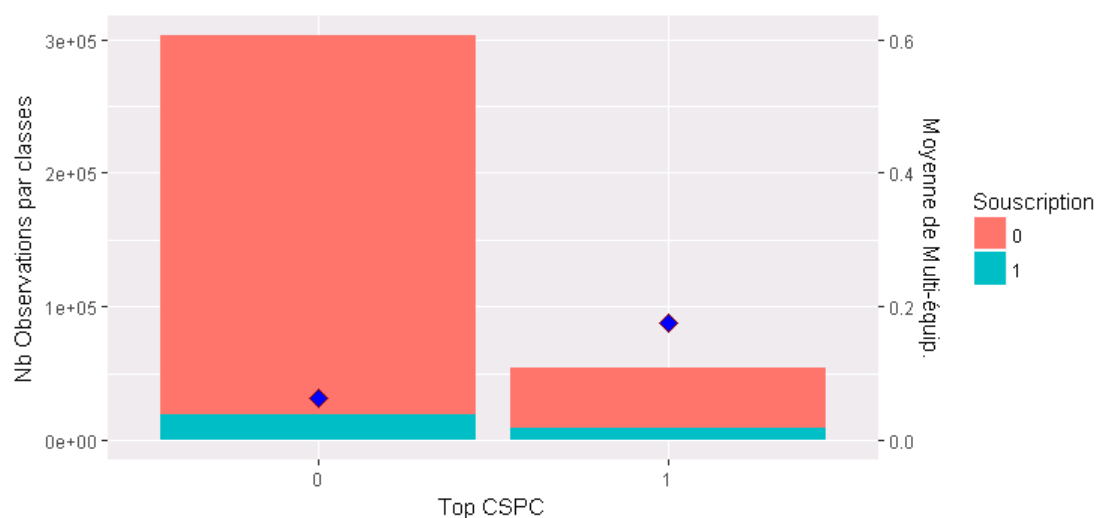


Figure 2.15 – Taux moyen de clients qui se multi-équipent en fonction de la variable *Indicatrice de la CSP*

Seulement un septième des clients ont leur CSP renseignée. Cela s’explique par le fait que la CSP n’est renseignée que lorsque celle-ci est un facteur de risque (c’est notamment le cas pour la *MRC*). Toutefois le taux est beaucoup plus élevé dans le cas où la CSP est renseignée. Cela peut s’expliquer par le fait que l’intermédiaire, connaissant la CSP, possède une meilleure connaissance des besoin du client et est plus apte à le multi-équiper.

- La **figure 2.16** représente la variable *Indicatrice de la banque du client* :
Cette variable vaut 1 si la banque du client est renseignée, sinon elle vaut 0.

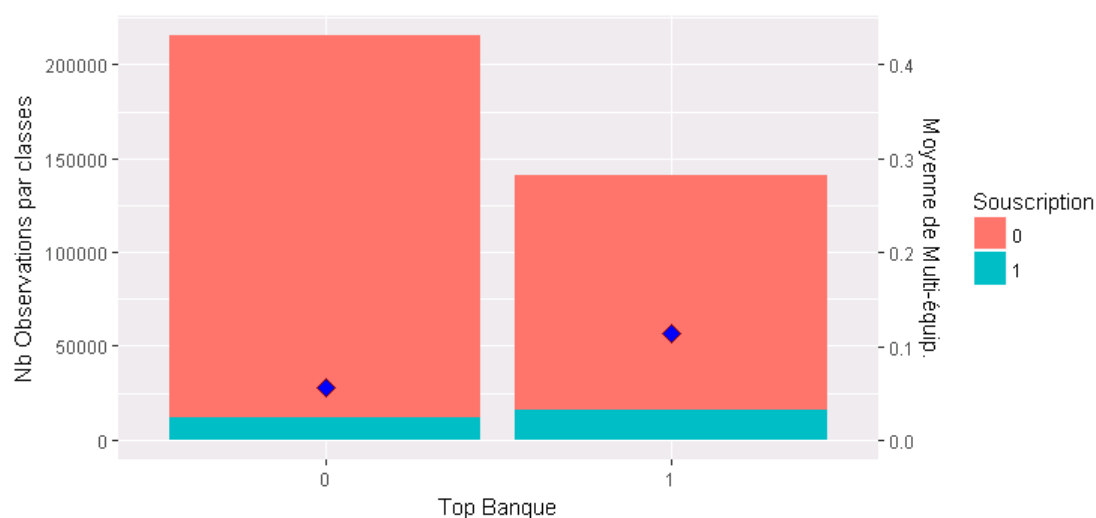


Figure 2.16 – Taux moyen de clients qui se multi-équipent en fonction de la variable *Indicatrice de la banque du client*

Les clients dont la banque est renseignée ont un probabilité deux fois plus élevée de se multi-équiper.

- La **figure 2.17** représente la variable *Indicatrice de la nullité de la prime IARD* :
Cette variable vaut 1 si la prime nette du client en produit *IARD* est supérieure à zéro, sinon elle vaut 0.

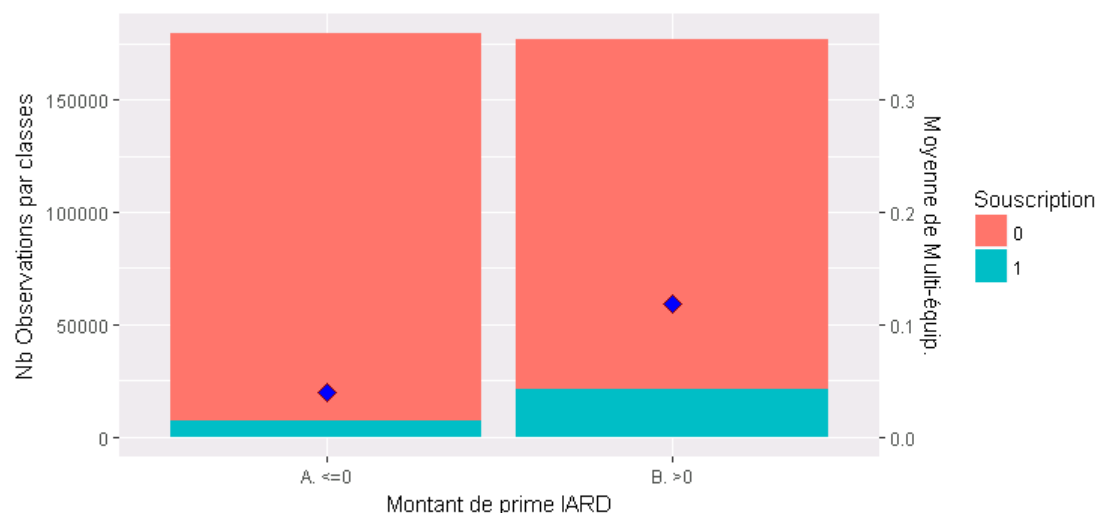


Figure 2.17 – Taux moyen de clients qui se multi-équipent en fonction de la variable *Indicatrice de la nullité de la prime IARD*

Ce graphique corrobore la conclusion de la **partie 2.2** selon laquelle les produits *IARD* sont des produits à potentiel.

- La **figure 2.18** représente la variable *Famille du produit d'appel (ou d'entrée)* :
Les produits d'appel ont été ici regroupés en des familles un peu plus générales. Toutefois les tendances qui avaient été identifiées restent les mêmes : les produits *Automobile* et *MRH* sont les produits avec un potentiel de multi-équipement fort.

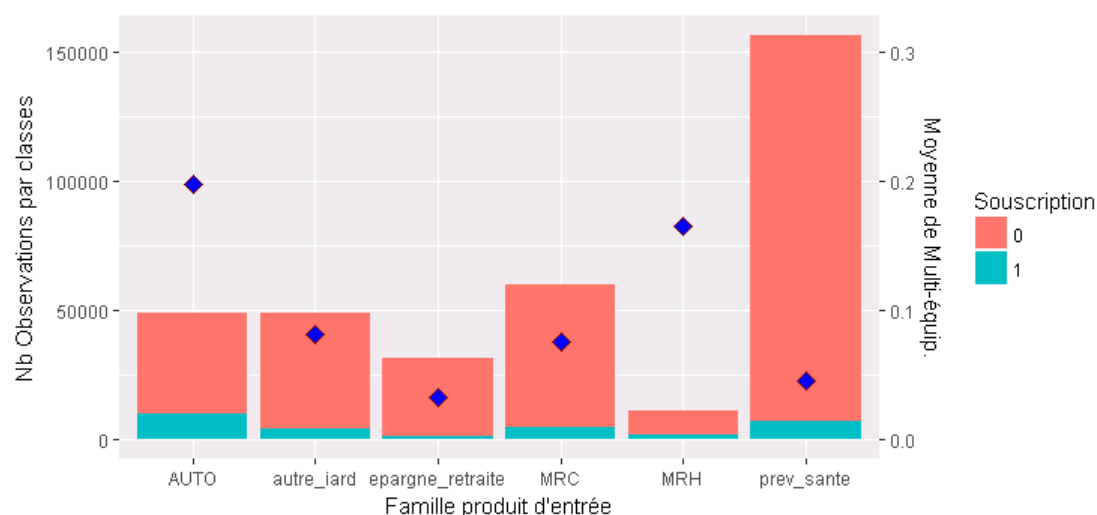


Figure 2.18 – Taux moyen de clients qui se multi-équipent en fonction de la variable *Famille du produit d'appel (ou d'entrée)*

- La **figure 2.19** représente la variable *Indicatrice de nombre d'enfants* :

Cette variable vaut 1 si le nombre d'enfants du client est supérieure à zéro, sinon elle vaut 0. Dans le cas où le nombre d'enfants n'est pas renseigné, le client est considéré sans enfant.

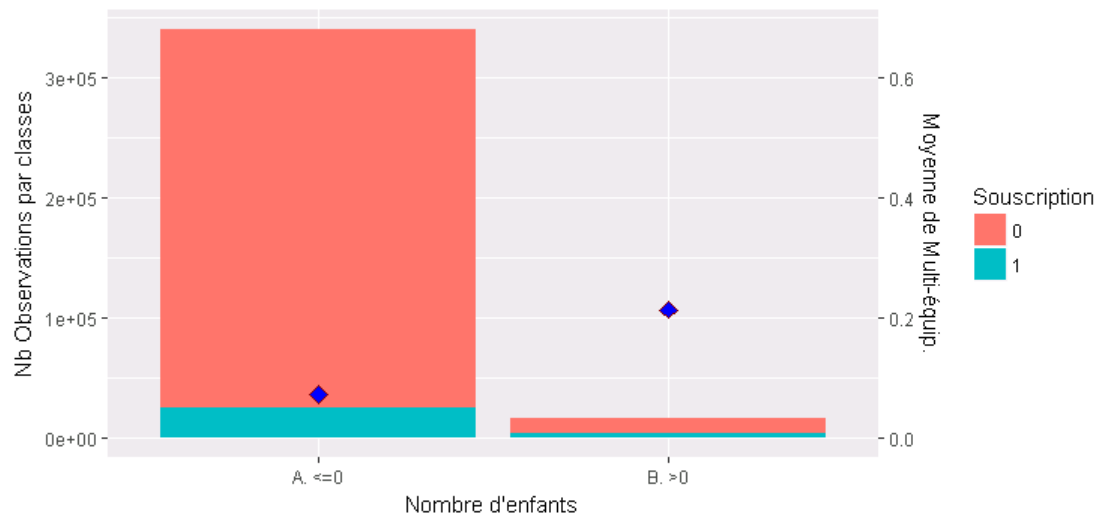


Figure 2.19 – Taux moyen de clients qui se multi-équipent en fonction de la variable *Indicatrice de nombre d'enfants*

Le fait d'avoir un enfant joue beaucoup sur le multi-équipement. Cela peut s'expliquer par le fait que le client s'équipe en produit pour ses enfants et de façon plus générale pour sa famille.

- La **figure 2.20** représente la variable *Nombre de contrats* :
Le nombre de contrats d'un client a été regroupé en trois catégories.

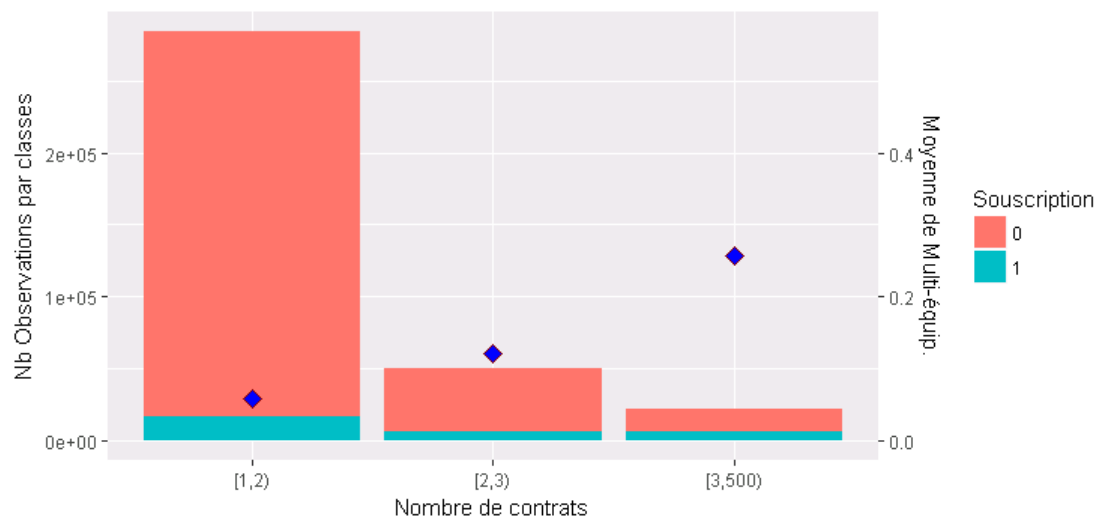


Figure 2.20 – Taux moyen de clients qui se multi-équipent en fonction de la variable *Nombre de contrats*

Le taux de multi-équipement augmente avec le nombre de contrats détenus par le client.

Remarque : la variable numérique *Nombre de contrats détenus* est laissé dans la base de données dans la mesure où des méthodes de sélection de variables seront réalisées.

- La **figure 2.21** représente la variable *Composition du panier produit* :

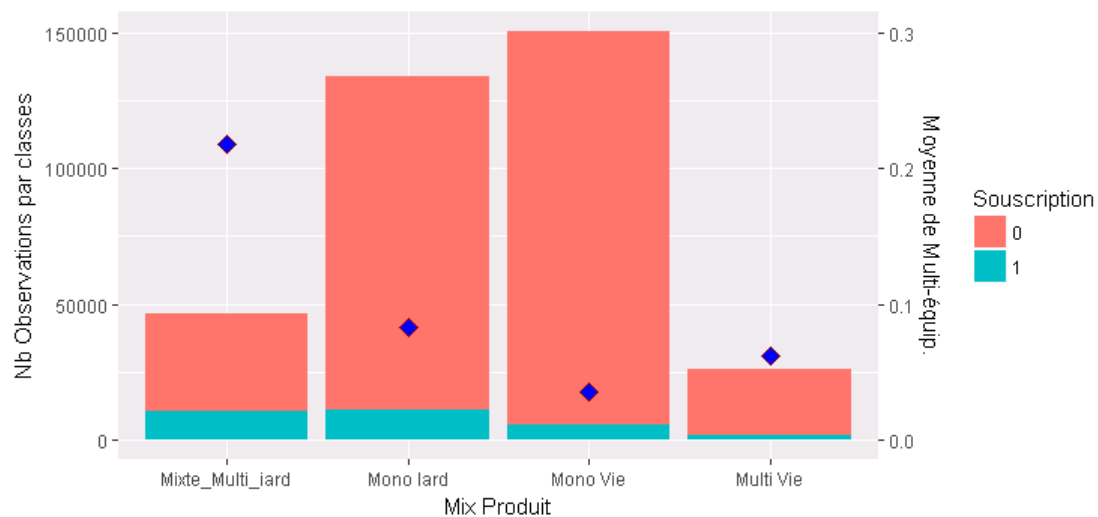


Figure 2.21 – Taux moyen de clients qui se multi-équipent en fonction de la variable *Composition du panier produit*

Détail des modalités de la variable :

- Mixte_Multi_IARD : regroupe les clients qui ont soit plusieurs contrats IARD au moins un contrat *IARD* et un contrat *non-IARD* ;
- Mono *IARD* : clients qui ne possèdent qu'un seul contrat *IARD* ;
- Mono *Vie* : clients qui ne possèdent qu'un seul contrat *Vie* ou un seul contrat d'assurance de *Personnes* ;
- Multi Vie : clients qui possèdent plusieurs contrats *Vie* et / ou d'assurance de *Personnes*.

Les clients uniquement équipés de produits vie ont une appétence très inférieure.

- La **figure 2.22** représente la variable *Délai entre la souscription du premier et du second contrat* :

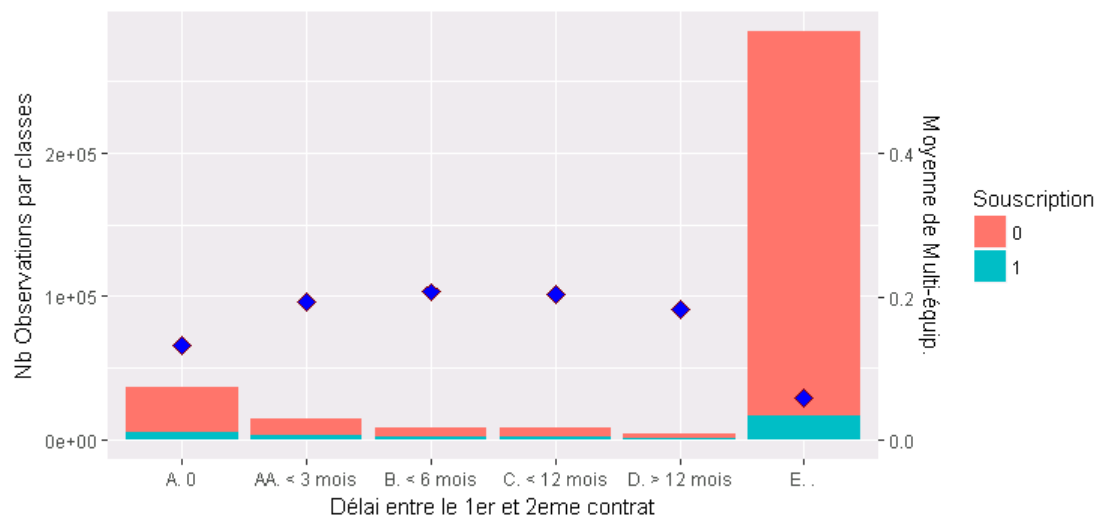


Figure 2.22 – Taux moyen de clients qui se multi-équipent en fonction de la variable *Délai entre la souscription du premier et du second contrat*

La modalité "A. 0" représente les clients qui ont souscrit leur second contrat en même temps que le premier. Le taux de multi-équipement de cette catégorie de clients est plus faible que celui des autres catégories car beaucoup de clients se multi-équipent le premier jour sans pour autant continuer à se multi-équiper par la suite.

La modalité "E." représente les clients qui n'ont pas souscrit de second contrat.

Il est décidé de regrouper les modalités "AA. < 3 mois", "B. < 6 mois", "C. < 12 mois" et "D. > 12 mois" car ces modalités présentent un taux similaire. De plus, cela permet d'opposer les clients qui se multi-équipent dès l'entrée en portefeuille des autres clients.

- La **figure 2.23** représente la variable *Délai entre la souscription du dernier et de l'avant-dernier contrat* :

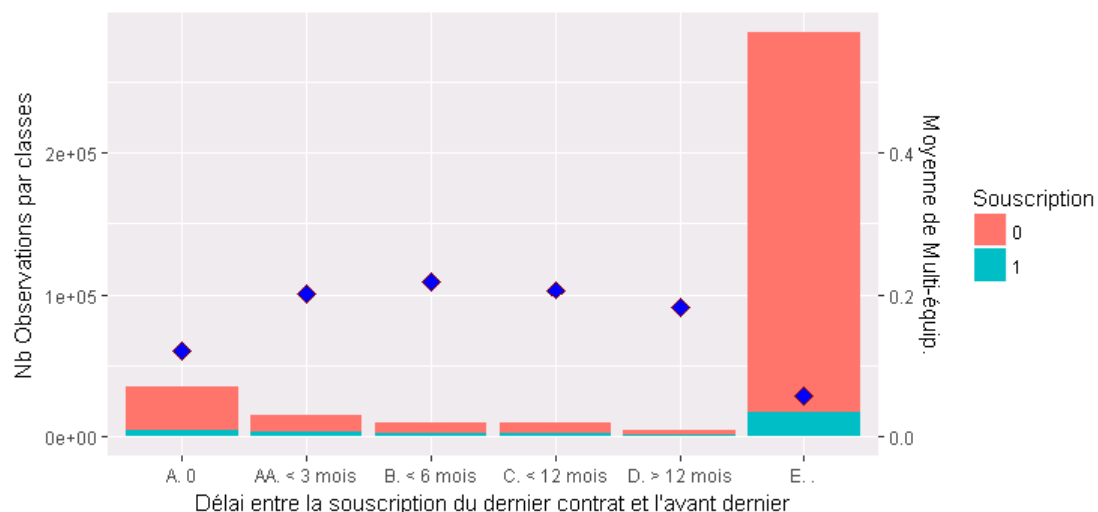


Figure 2.23 – Taux moyen de clients qui se multi-équipent en fonction de la variable *Délai entre la souscription du dernier et de l'avant-dernier contrat*

La remarque qui a été faite pour la modalité "A. 0" du **graphique 2.21** peut être réitérée

pour ce graphique dans la mesure où une grande partie des clients multi-équipés possèdent deux contrats.

La modalité "E." représente les clients qui n'ont pas souscrit d'avant-dernier contrat, ces clients sont donc les mêmes que ceux qui n'ont pas souscrits de second contrat.

Il est décidé de regrouper les modalités "AA. <3 mois", "B. < 6 mois", "C. < 12 mois" et "D. > 12 mois" car ces modalités présentent un taux similaire.

2.4 Construction de la base pour la rentabilité moyenne des contrats

Dans cette partie est présentée la construction de la base de données permettant la prédiction de la valeur moyenne en cas de souscription.

2.4.1 Construction de la base d'apprentissage

L'objectif est de déterminer la rentabilité moyenne des contrats (i.e. : la valeur moyenne) que pourraient souscrire les assurés présents en portefeuille.

Ainsi, nous ne nous intéressons qu'à la valeur des affaires nouvelles. De plus, nous ne considérons que les affaires nouvelles réalisées en 2018. Ce choix permet d'avoir une valeur moyenne la plus proche possible du tarif en vigueur.

La base de données est donc constituée de toutes les affaires nouvelles qui ont été réalisées lors de l'année 2018 et dont la valeur du contrat a été calculée.

Toutefois, c'est la valeur hors offres commerciales moyenne du contrat que nous chercherons à déterminer. Ainsi, la valeur du contrat en portefeuille, qui prend en compte les offres commerciales, n'est pas directement utilisable. Il convient alors de corriger les primes en supprimant les offres commerciales accordées.

Une fois les primes hors offres commerciales calculées, la projection des flux est à nouveau réalisée pour aboutir à la valeur hors offres commerciales des contrats.

Enfin, les variables explicatives, qui sont utilisées pour calculer les valeurs moyennes, sont :

- les zoniers *Automobile* et *MRH* ;
- le réseau de distribution, qui est séparé en deux modalités : réseau "agent" et réseau "autre" (qui est principalement composé de courtiers et est donc noté "courtier") ;
- la nature du client, qui est séparée en trois modalités : Entreprise, Petite Entreprise et PRO ;
- l'âge du client qui est parfois renseigné pour les professionnels.

L'âge est découpé en quartile. Une cinquième classe est créée pour les clients dont l'âge n'est pas renseigné.

Une base de données est créée pour chaque produit. Ainsi trois bases de données sont construites, la première pour les produits *Automobile hors Engins Spéciaux* (289 000 observations), la seconde pour les produits *Engins Spéciaux* (48 000 observations) et la dernière pour les produits *MRH & PNO* (117 000 observations) (le choix de ces produits est expliqué dans la **partie 4.1.1**).

Enfin, le découpage en quartile de la variable "Age" est amené à varier d'une base à l'autre, la distribution n'étant pas forcément la même.

Remarque : la base qui est ici construite a vocation à modéliser la valeur des contrats que le clients va souscrire. Or, dans la mesure où de tels contrats ne sont pas encore souscrits, il n'est pas possible de connaître le risque associé au client et donc sa valeur. Les variables qui sont utilisées sont assez peu explicatives. C'est pourquoi la modélisation de la distribution de la valeur sera aussi effectuée dans la suite.

2.4.2 Statistiques sur la variable réponse

Dans cette partie nous observons les statistiques d'ordre de la variable réponse dans chacune des trois bases de données.

1. Statistiques sur la valeur des contrats dans la base contenant les produits Automobile hors Engins Spéciaux :

Minimum	Premier Quartile	Médiane	Moyenne	Troisième Quartile	Maximum
-3 225	206	335	418	538	6 798

La distribution de la valeur des contrats Automobile hors Engins Spéciaux présente des valeurs extrêmes à gauche et à droite.

2. Statistiques sur la valeur des contrats dans la base contenant les produits Engins Spéciaux :

Minimum	Premier Quartile	Médiane	Moyenne	Troisième Quartile	Maximum
0	93	178	236	271	4 946

La valeur des contrats Engins Spéciaux est très peu dispersée. En effet 75 % de la distribution se situent entre les valeurs 93 et 271.

3. Statistiques sur la valeur des contrats dans la base contenant les produits MRH & PNO :

Minimum	Premier Quartile	Médiane	Moyenne	Troisième Quartile	Maximum
16	105	182	275	364	17 400

La distribution de la valeur des contrats MRH & PNO présente des valeurs extrêmes à droite.

Pour chacune des bases de données, la distribution de la valeur des contrats présente des valeurs aberrantes. Ainsi, il est décidé de supprimer les valeurs au-delà du quantile à 99 % de la valeur absolue des variables réponses.

Dès lors :

- pour les contrats Automobile hors Engins Spéciaux, les valeurs gardées sont comprises entre - 1 549 et 1 549 ;
- pour les contrats Engins Spéciaux, les valeurs gardées sont comprises entre - 1 333 et 1 333 ;
- pour les contrats MRH & PNO, les valeurs gardées sont comprises entre - 1 130 et 1 130.

2.4.3 Statistiques sur les variables explicatives

La variable qui est modélisée, la valeur du contrat, est continue. Ainsi dans cette partie, les diagrammes en boîte en fonction des modalités des variables explicatives sont représentés.

La largeur d'une boîte est proportionnelle au nombre d'individus dans la modalité. Une boîte large contient un nombre important d'individus, inversement une boîte mince présente peu d'individus.

Dans chaque boîte est présent un point qui correspond à la moyenne. Une moyenne éloignée de la médiane indique une forte asymétrie.

1. Diagrammes en boîte pour la base des produits **Automobile hors Engins Spéciaux** :

- La **figure 2.24** représente la répartition de la valeur des contrats en fonction des variables *Nature du client*, *Age* et *Réseau de distribution*.

Bien que les distributions soient proches pour les variables *Nature du client* et *Réseau de distribution*, des différences de valeur se distinguent malgré tout en moyenne.

Concernant la variable *Age*, le diagramme en boîte nous interroge sur la pertinence de séparer l'âge en quatre modalités lors que celui-ci est renseigné.

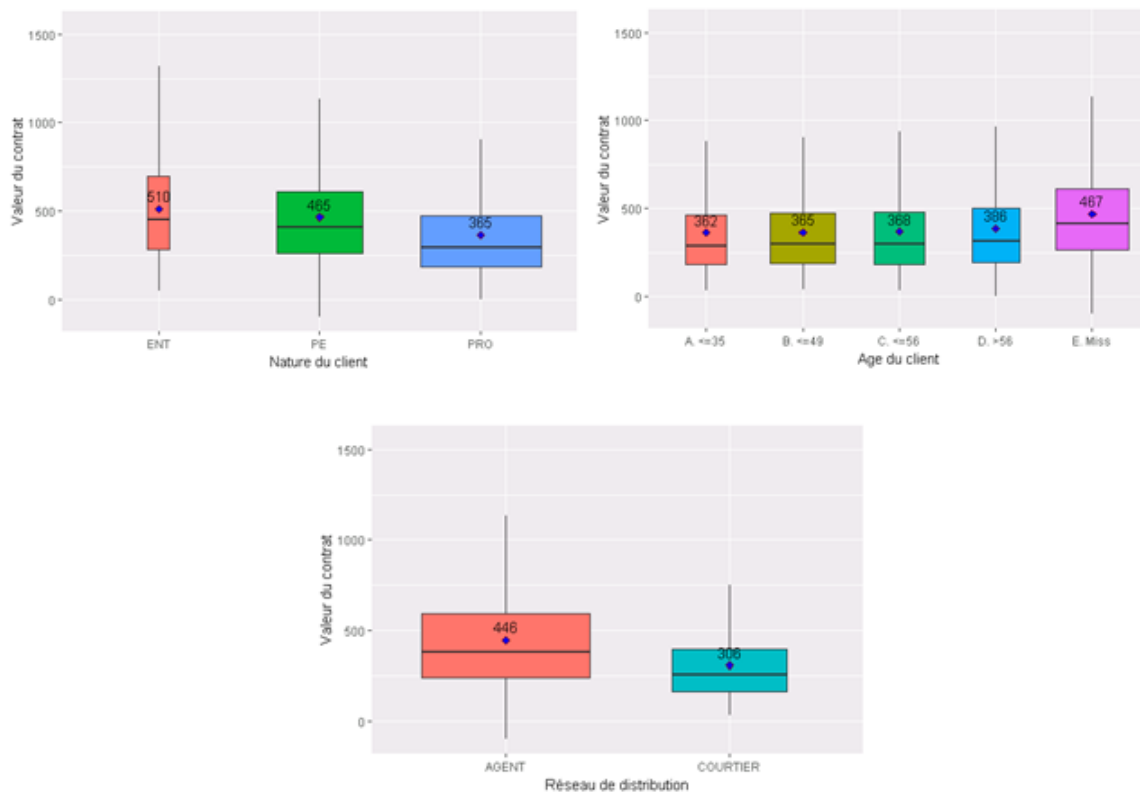


Figure 2.24 – Diagrammes en boîte de la valeur des contrats en fonction des variables *Nature du client*, *Age* et *Réseau de distribution*

- La **figure 2.25** représente la répartition de la valeur des contrats en fonction des variables *Zonier responsabilité civile (RC)*, *Zonier bris de glace (BDG)*, *Zonier vol* et *Zonier dommage (DOM)*.

La tendance est légèrement croissante pour les *Zoniers RC*, *Vol* et *Dommages*. Concernant le *Zonier bris de glace*, la tendance est décroissante pour les premières zones puis stagne.

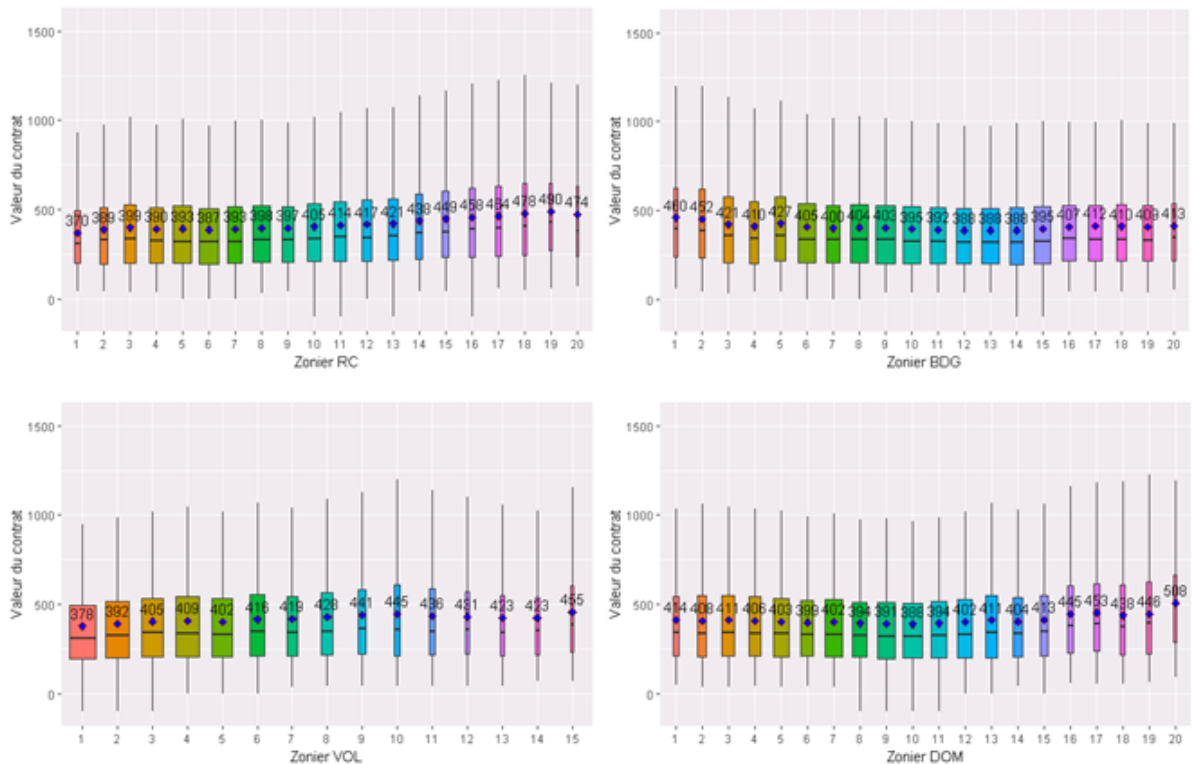


Figure 2.25 – Diagrammes en boîte de la valeur des contrats en fonction des variables *Zoniers*

2. Diagrammes en boîte pour la base des produits **Engins Spéciaux** :

- La **figure 2.26** représente la répartition de la valeur des contrats *Engins Spéciaux* pour les variables *Réseau de distribution* et *Nature du client*.

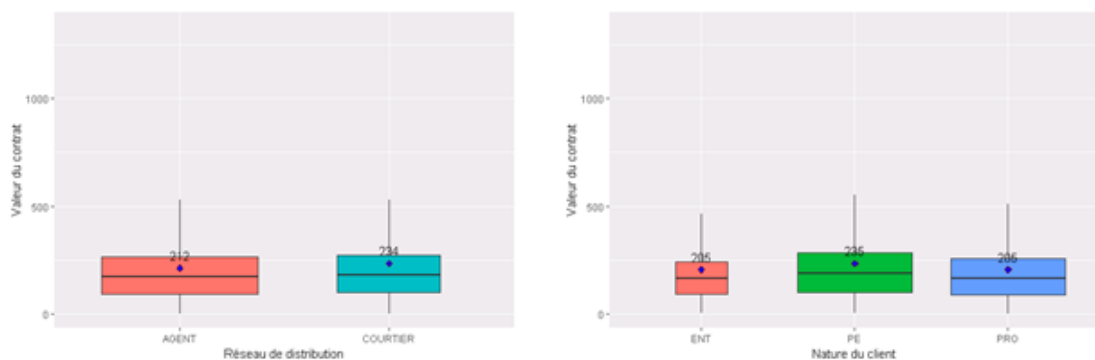


Figure 2.26 – Diagrammes en boîte de la valeur des contrats en fonction des variables *Réseau de distribution* et *Nature de client*

Les variables explicatives sont certes différentes en moyenne, mais les distributions de la valeur au sein de chaque modalité sont quasiment identiques. Les variables sont d'une part peu informatives et d'autre part la variable réponse est très peu dispersée.

3. Diagrammes en boîte pour la base des produits **MRH & PNO** :

- La **figure 2.27** représente la répartition de la valeur des contrats *MRH* & *PNO* pour les variables *Nature de client*, *Age* et *Réseau de distribution*.

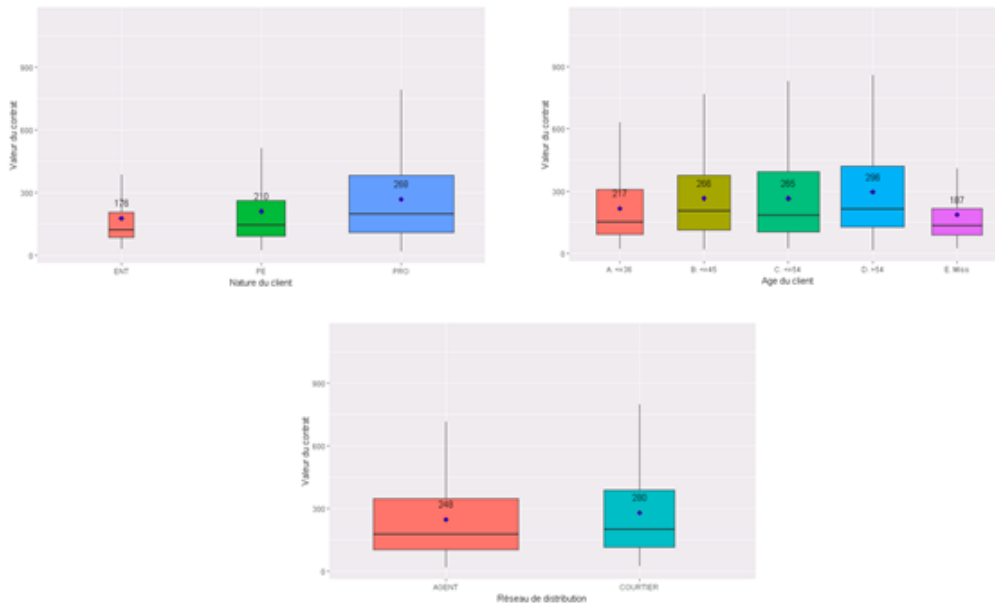


Figure 2.27 – Diagrammes en boîte de la valeur des contrats en fonction des variables *Nature de client*, *Age* et *Réseau de distribution*

- La **figure 2.28** représente la répartition de la valeur des contrats *MRH* & *PNO* pour les variables *Zonier tempête grêle neige (TGN)* et *Zonier vol*.

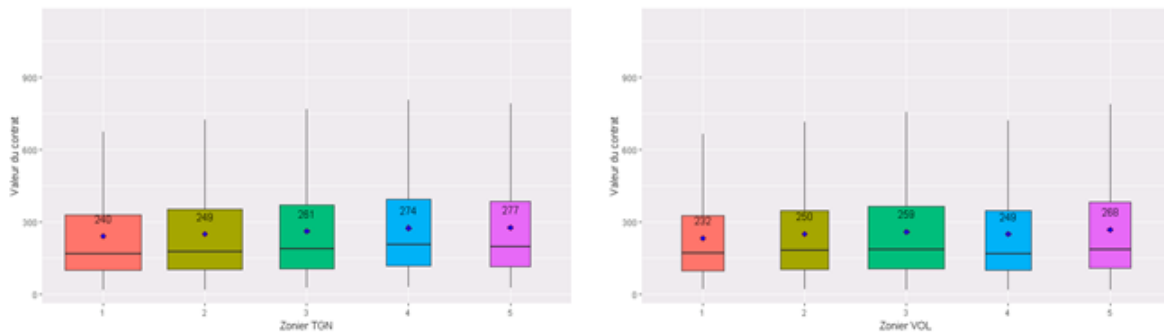


Figure 2.28 – Diagrammes en boîte de la valeur des contrats en fonction des variables *Zoniers*

Les remarques qui ont été faites pour les deux produits précédents peuvent être ici réitérées. Les distribution entre les différentes modalités sont très proches.

L'observation des différents diagrammes en boîte pour chacun des produits nous informe du caractère faiblement informatif des variables explicatives. C'est pourquoi **la distribution de la valeur du contrat sera aussi modélisée, afin de prendre en compte l'aléa sur la prédiction de la valeur en cas de souscription.**

Chapitre 3

Modéliser la probabilité de souscription d'un nouveau contrat à horizon d'une année

L'objectif de cette partie est de modéliser la probabilité qu'a un client de souscrire un nouveau contrat durant l'année. Cette probabilité sera calculée pour tous les clients professionnels du portefeuille de moins de deux ans d'ancienneté. L'objectif est donc de résoudre un problème de classification binaire ; pour se faire nous allons recourir à deux méthodes d'apprentissage : la régression logistique (**partie 3.1**) et le gradient boosting (**partie 3.2**). La base de données de 2018 permettra de vérifier que les prédictions du modèle restent stables dans le temps. Ensuite, les modèles statistiques seront appliqués (**partie 3.3**) et comparés entre eux (**partie 3.4**).

3.1 La régression logistique

3.1.1 La famille exponentielle

La famille exponentielle regroupe toutes les lois de probabilité dont la fonction de densité, par rapport à la mesure dominante appropriée, s'écrit :

$$f(y|\theta, \phi) = \exp \left(\frac{y \cdot \theta - b(\theta)}{a(\phi)} + c(y, \phi) \right)$$

θ est appelé le paramètre naturel de la loi, il s'agit du paramètre d'intérêt de la modélisation ; ϕ est appelé le paramètre de nuisance de la loi.

Détaillons quelques propriétés :

Soit Y une variable aléatoire dont la loi appartient à la famille exponentielle, alors :

$$E[Y] = b'(\theta) \text{ et } Var(Y) = b''(\theta) * \phi$$

L'espérance de la variable Y est donc liée au paramètre θ via la fonction b' .

La fonction canonique g d'une loi de la famille exponentielle est définie telle que :

$$g(\mu) = \theta \text{ où } \mu \text{ est l'espérance de la loi}$$

$$\text{Ainsi } g = b'^{-1}$$

La loi de Bernoulli appartient à la famille exponentielle, sa densité valant :

$$f(y|\theta, \phi) = p^y \times (1-p)^{1-y} = \exp \left(y \log \frac{p}{1-p} - \log \frac{1}{1-p} \right) = \exp (y\theta - b(\theta))$$

$$\text{Où } \theta = \log \left(\frac{p}{1-p} \right)$$

Et sa fonction de lien canonique est la fonction *logit* :

$$\text{logit}(x) = \log \left(\frac{p}{1-p} \right)$$

3.1.2 Le modèle linéaire gaussien

Nous définissons les quantités suivantes :

- Y une variable aléatoire réelle, appelée variable à expliquer. La variance de Y est supposée constante, il s'agit de l'hypothèse d'homoscédasticité ;
- $(X_1 | \dots | X_k) = X$ sont les k variables explicatives, avec $X_1 = (1 \dots 1)$;
- $\beta_1 \dots \beta_k$ sont les k paramètres du modèle qui sont à estimer ;
- E est un vecteur correspondant à k réalisations indépendantes d'un bruit gaussien ϵ centré et de variance inconnue notée σ .

Le modèle s'écrit :

$$Y = \sum_{i=1}^k (X \times \beta) + E$$

Les paramètres sont déterminés par la minimisation des moindres carrés :

$$\hat{\beta} = \operatorname{argmin}_{\beta} \|Y - X\beta\|^2$$

Les hypothèses du modèle linéaire sont :

- X doit être de rang plein et doit être non aléatoire ;
- les erreurs doivent être indépendantes, centrées et de variance constante ;
- le nombre d'observations doit être supérieur au nombre de paramètres.

Propriétés sur les estimateurs $\hat{\beta}$ et $\hat{\sigma}$:

- L'estimateur de β suit une loi normale : $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X'X)^{-1})$
- $(n - k) \times \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-k}^2$

3.1.3 Le modèle linéaire généralisé (GLM)

Le modèle linéaire qui vient d'être présenté peut être étendu à toute variable aléatoire suivant une loi appartenant à la famille exponentielle. Dans un tel modèle, $\mathbf{Y}|\mathbf{X}$ suit alors une distribution appartenant à cette famille. L'hypothèse sur l'indépendance conditionnelle des résidus est la même que dans le modèle linéaire.

Le modèle GLM fait intervenir une fonction de lien entre $\mathbf{E}[\mathbf{Y}|\mathbf{X}]$ et $\mathbf{X} \times \boldsymbol{\beta}$ tels que :

$$g(E[Y|X]) = X \times \beta$$

Un choix possible de g est la fonction de lien canonique de la loi.

Dans notre modélisation, la variable réponse est une variable binaire et sa distribution associée correspond à une distribution de Bernoulli, dès lors notre choix de g se porte sur la fonction de lien canonique : la fonction *logit*.

La régression logistique est le nom donné au GLM ayant une variable réponse suivant une loi de Bernoulli et ayant pour fonction de lien la fonction *logit*.

Remarques :

- Le choix de la fonction g constitue une hypothèse du modèle. Il est possible de choisir n'importe quelle fonction bijective de $[0,1]$ vers $\mathbb{R}$ pour g (par exemple la fonction *probit* ou *log-log*).
- Le modèle linéaire gaussien est un cas particulier du GLM, la distribution de la famille exponentielle est la distribution normale et la fonction de lien est l'identité.

En résumé, un modèle GLM est composé de trois éléments :

- la variable à expliquer, qui appartient à une **famille exponentielle** ;
- les variables explicatives, qui ne sont pas colinéaires (i.e. : **$\mathbf{X}$ est de rang plein**) ;
- une **fonction de lien**.

3.1.4 Estimation des paramètres

Les paramètres (β_k) du modèle sont déterminés par maximum de vraisemblance.

Ainsi, la vraisemblance s'écrit :

$$L(\theta, \phi|Y) = \prod_{i=1}^n \exp \left(\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right) = \exp \left(\sum_{i=1}^n \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right)$$

Et la log-vraisemblance s'écrit :

$$\log(L(\theta_i|y_1 \dots y_n)) = \sum_{i=1}^n \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)$$

Les paramètres (β_k) qui maximisent cette quantité vérifient la condition du première ordre.

En notant :

$$\mu_i = E[Y_i] \text{ et } \eta_i = g(\mu_i) = X_i\beta;$$

nous avons pour la i -ème observation et le j -ème paramètre :

$$\frac{\partial \log(L_i)}{\partial \beta_j} = \frac{\partial \log(L_i)}{\partial \mu_i} \times \frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \times \frac{y_i - \mu_i}{V(Y_i)} X_{i,j}$$

Ainsi, il faut résoudre :

$$\forall j : \sum_{i=1}^n \frac{\partial \log(L_i)}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial \mu_i}{\partial \eta_i} \times \frac{y_i - \mu_i}{V(Y_i)} X_{i,j} = 0$$

Ces équations ne peuvent être résolues de façon analytique, toutefois elles peuvent être résolues numériquement via une descente du gradient.

3.1.5 Pénalisation Lasso et Ridge

Dans un modèle GLM, la log-vraisemblance négative correspond à la fonction de coût à minimiser. Toutefois, il est possible d'ajouter au coût une fonction de régularisation (ou pénalisation) qui permet d'ajouter une information a priori :

$$\text{Coût} = -\log(L) + g(\beta)$$

La fonction g reflète l'a priori que l'on possède sur le paramètre à estimer.

Les a prioris les plus répandus sont la parcimonie du modèle (seul un petit nombre de paramètres sont non nuls) ou l'a priori selon lequel l'amplitude des paramètres est contenue.

3.1.5.1 PÉNALISATION LASSO

La pénalisation Lasso (*Least Absolute Shrinkage and Selection Operator*) correspond à la formalisation de l'a priori de parcimonie sur les paramètres. Elle consiste en l'ajout d'une contrainte de norme L_1 sur les paramètres :

$$\text{Coût} = -\log(L) + \lambda * |\beta|_1$$

λ correspond à l'importance qui est accordée à la régulation et donc à l'a priori.

- Si λ tend vers 0 le coût correspond à celui d'un modèle GLM classique sans l'hypothèse de parcimonie.
- Si λ tend vers $+\infty$ alors, mécaniquement, quelque soit le vecteur β différent du vecteur nul :

$$\lambda * |\beta|_1 \gg -\log(L)$$

et donc le coût est minimisé pour $\beta = \vec{0}$. Cela correspond au modèle sans paramètre qui est le modèle le plus parcimonieux possible.

Cette contrainte va contracter la valeur des paramètres. De plus, la pénalité va permettre à certains coefficients de valoir exactement zéro.

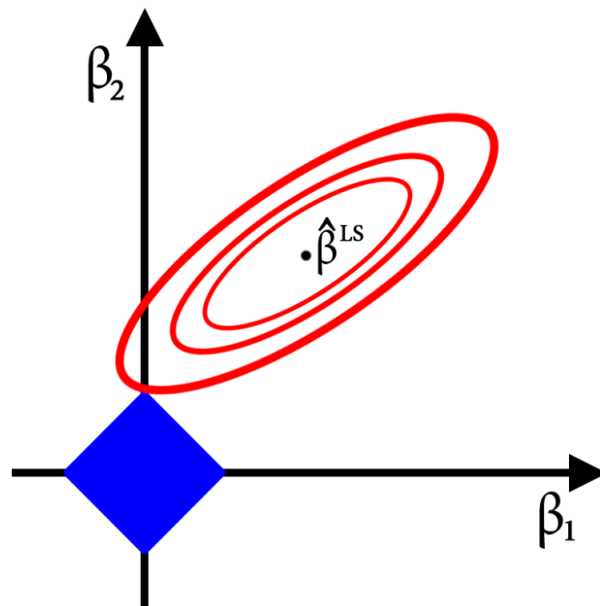


Figure 3.1 – **Contrainte Lasso en bleu et fonction de Coût en rouge**

Source : Tibshirani Robert, Hastie Trevor et Wainwright Martin, *Statistical Learning with Sparsity : The Lasso and Generalizations*, Chapman and Hall, 2015

La **figure 3.1** montre que la minimisation de la contrainte se fait sur l'un des bords du losange bleu, ce qui correspond à la nullité de l'un des coefficients.

La pénalisation Lasso est donc efficace pour sélectionner des variables.

Il est bon de noter que cette méthode sélectionne aussi certaines modalités des variables qualitatives et pas obligatoirement toutes les modalités comme cela peut être le cas d'autres méthodes de sélection (Step by Step par exemple).

En effet, dans les modèles GLM, les variables qualitatives sont "*one-hot* encodées", cela signifie qu'une variable à k modalités est transformée en k variables, la j -ième variable vaut 1 si l'individu possède la j -ième modalité, 0 sinon.

Une variable est ensuite prise comme variable de référence pour que le modèle soit identifiable (pour rappel, la matrice X doit être de rang plein).

Ainsi le modèle doit estimer $k-1$ paramètres s'il y a k modalités. La sélection Lasso regardant les paramètres de façon indépendante, il est dès lors possible que certaines modalités soit activées (le coefficient associé est non nul) alors que d'autres ne le sont pas.

3.1.5.2 PÉNALISATION RIDGE

La pénalisation Ridge correspond à l'ajout d'une contrainte de norme L_2 sur les paramètres :

$$\text{Coût} = -\log(L) + \lambda * |\beta|_2$$

Comme dans le cas de la régularisation Lasso, λ correspond à l'importance que l'on accorde à la pénalisation L_2 . Contrairement à la pénalisation Lasso, la pénalisation Ridge ne va pas attribuer la valeur nulle à certains coefficients mais va modérer l'amplitude globale des coefficients. Cette pénalisation correspond à l'a priori selon lequel les paramètres ne prennent pas de valeurs trop importantes.

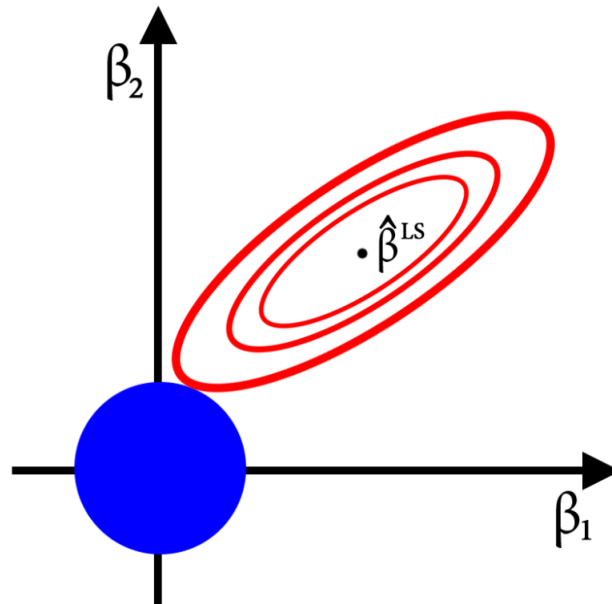


Figure 3.2 – **Contrainte Ridge en bleu et fonction de Coût en rouge**

Source : Tibshirani Robert, Hastie Trevor et Wainwright Martin, *Statistical Learning with Sparsity : The Lasso and Generalizations*, Chapman and Hall, 2015

3.1.6 Les limites du modèle de régression logistique

Plusieurs critiques peuvent être faites à l'encontre des modèles linéaires. La première concerne le caractère linéaire, puisqu'il impose que :

$$g(E[Y|X = x]) = a \cdot x, \text{ dans le cas où il n'y a qu'une seule variable explicative.}$$

Ainsi, les autres types de relation entre X et Y ne peuvent être captées par le modèle, sauf si le statisticien les a préalablement identifiées. Dans ce cas, il est possible de créer de nouvelles variables qui captent mieux le phénomène identifié (découpage en classe, transformation en polynômes, utilisation de splines).

Par extension, cela signifie, sous l'hypothèse de linéarité, que les Y sont linéairement séparables dans le nuage de points X , en d'autres termes qu'il existe un hyperplan sépareur. Or, cela n'est souvent pas le cas.

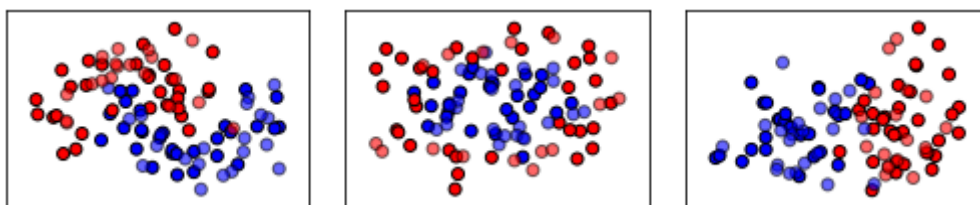


Figure 3.3 – Types de structures de données

Par exemple, les deux premiers nuages de points de la **figure 3.3** ne sont pas linéairement séparables.

Une seconde limite est le fait que le modèle linéaire ne prend pas en compte les interactions potentielles entre les variables explicatives. Cela reste à la charge du statisticien d'identifier de potentiels liens et de croiser ensuite ces variables dans le modèle. En général, seul un nombre restreint d'interactions est testé, le plus souvent entre les variables les plus impactantes. La connaissance de ces interactions fait généralement appel à des connaissances métiers et qui sont fournies par des experts qui ne sont pas nécessairement statisticiens.

3.2 Le gradient boosting

3.2.1 Pourquoi un tel modèle ?

Comme nous l'avons vu dans la partie précédente, il n'existe pas nécessairement de relation linéaire entre les variables explicatives et la variable à expliquer. De plus, les interactions entre les variables ne sont pas testées par le modèle de régression logistique et sont laissées à la charge du statisticien.

Ce sont deux inconvénients qui ne sont pas présents dans les méthodes empiriques que sont les arbres de décision. C'est pourquoi de tels algorithmes d'apprentissage seront entraînés sur les données en parallèle de la régression logistique. Dans la suite, nous nous concentrerons uniquement sur les arbres de classification et de régression (CART : *Classification And Regression Trees*) qui sont les arbres les plus utilisés.

3.2.2 Les arbres de classification et de régression

3.2.2.1 DÉFINITION

Un arbre de décision est constitué de partitions successives du nuage de points. Nous appellerons de telles partitions des "*splits*". Chaque partition est binaire et porte sur une seule variable et une seule valeur seuil à la fois. Ainsi, à chaque nœud de l'arbre, l'ensemble de points est divisé en deux sous-ensembles. Si aucune partition ne se produit, il s'agit d'un nœud terminal.

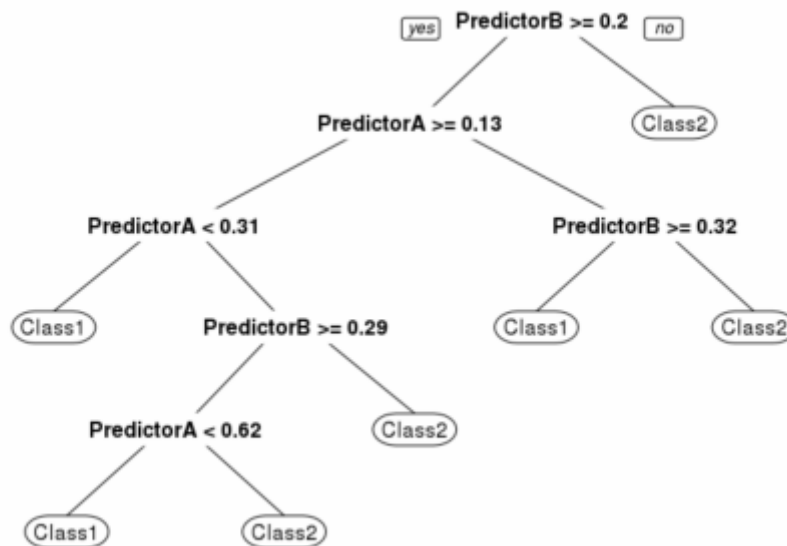


Figure 3.4 – **Arbre de classification**

Source : Boyer Claire, « Apprentissage statistique », ISUP, 2018

Dans l'exemple de la **figure 3.4** ci-dessus, l'individu ayant pour coordonnées (**A=0.2**, **B=0.5**) sera classé par l'arbre comme appartenant à la classe 2.

3.2.2.2 CONSTRUCTION D'UN ARBRE CART

1. Idée liminaire

Pour aboutir à un tel arbre, il faut partir initialement de l'ensemble du nuage de points X . Puis l'espace est divisé selon une variable explicative et un certain seuil. Chaque sous-espace est ensuite à nouveau divisé selon une variable et un certain seuil. Il peut s'agir d'une variable et d'un seuil différents pour chaque sous-espace. Le processus se poursuit jusqu'à ce qu'une condition d'arrêt soit rencontrée.

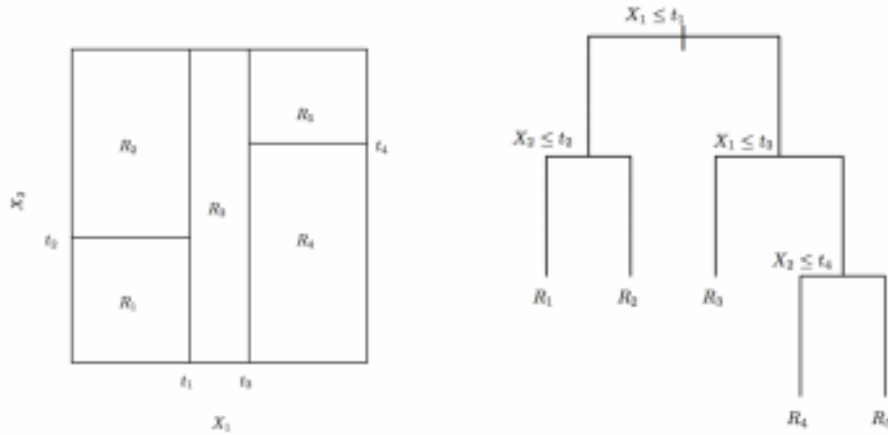


Figure 3.5 – **Exemple : ensemble de données divisé en sous-espace et arbre de décision associé**

Source : Claire Boyer Apprentissage Statistique Sorbonne Université 2018

Cette méthode permet d'aboutir à une partition de l'espace en hypercubes. De plus, la méthode de construction des arbres est dite sans regret. En effet, une fois la partition choisie, l'algorithme traite les sous-espaces générés de façon indépendante.

2. Construction des *splits*

Intéressons nous maintenant à la façon dont l'algorithme choisit la partition à effectuer. Le but est d'obtenir les sous-espaces les plus homogènes possibles, ainsi il faut quantifier cette homogénéité pour chacun des *splits* possibles et choisir alors celui qui la maximise.

En notant N l'ensemble de points dont il faut mesurer l'homogénéité :

- si la variable à prédire est quantitative, la principale mesure de l'homogénéité est la variance :

$$V(N) = \sum_{i: x_i \in N} (y_i - \bar{y}_N)^2;$$

- si la variable à prédire est qualitative et possède K modalités, il faut d'abord calculer la distribution des modalités :

$$p_N = (p_{N,1} \dots p_{N,K}), \text{ où } \forall k \in \{1 \dots K\} : p_{N,k} = \frac{\#\{i : x_i \in N, y_i = k\}}{|N|};$$

il existe ensuite deux principales mesures de l'homogénéité :

- l'index de Gini :

$$V(N) = \sum_{i=1}^K (p_{N,i})(1 - p_{N,i})$$

- l'entropie :

$$V(N) = - \sum_{i=1}^K p_{N,i} \log_2(p_{N,i})$$

Maintenant que l'homogénéité d'un ensemble de points est définie, nous introduisons le gain d'information pour un *split*, soit :

- I la mesure d'homogénéité qui correspond à la variance, l'indice de Gini ou l'entropie ;
- j la variable utilisée pour le *split* ;
- t la valeur seuil utilisée pour le *split* ;
- N_L, N_R les nouveaux nuages de points résultant du *split*.

$$Gain(j, t) = I(N) - \frac{|N_L(j, t)|}{|N|} * I(N_L(j, t)) - \frac{|N_R(j, t)|}{|N|} * I(N_R(j, t)) .$$

Ainsi, la variable $\hat{j}$ et le seuil $\hat{t}$ correspondent à : $\hat{j}, \hat{t} = \operatorname{argmin}_{j, t} Gain(j, t)$

3. Conditions d'arrêt de *splits*

Les principales conditions d'arrêt sont :

- la profondeur maximale de l'arbre ;
- le noeud terminal ne contient qu'un individu ;
- l'homogénéité du sous-ensemble de point dépasse un certain seuil ;
- l'erreur de test augmente.

La paramétrisation des conditions d'arrêt reste à la charge du statisticien, elle constitue une étape importante du modèle et est très influente sur ses performances.

En effet si elles sont trop restrictives, l'arbre aura tendance à être très vaste et donc à sur-apprendre. Dans ce cas, la variance des prédictions sera très importante.

Si elles ne sont pas assez restrictives, l'arbre risquera d'être peu profond et possédera un biais important.

3.2.3 Le boosting et le gradient boosting

3.2.3.1 LE BOOSTING

Le boosting est une méthode ensembliste, c'est-à-dire qu'elle agrège plusieurs modèles, parfois contradictoires, pour aboutir à une prédiction.

Le but du boosting est de combiner linéairement des modèles faibles d'apprentissage (ou *Weak Learners*) pour aboutir à un modèle robuste et efficace (*Strong Learner*).

Le modèle combiné $g_B(x)$ vaut :

$$\sum_{i=1}^B \eta_i * h_i(x) \text{ avec } \eta_i \in \mathbb{R} \text{ et } h_i \text{ est un } \textit{Weak Learner}$$

Un *Weak Learner* est un modèle très simple, dont les performances sont légèrement meilleures que le hasard.

Par exemple un *Weak Learner* peut être un arbre de décision de profondeur 1 ou un modèle linéaire avec une ou deux variables explicatives.

Ainsi, le boosting consiste à résoudre le problème d'optimisation suivant :

$$(\hat{\eta}_i, \hat{h}_i)_{i \in 1:B} = \operatorname{argmin} \frac{1}{n} \sum_{i=1}^n l(y_i, g_B(x_i))$$

où l est la fonction de perte associée au problème.

Toutefois, la minimisation d'une telle fonction est très difficile. En effet, le problème consiste en la détermination de B estimateurs, où souvent B correspond à plusieurs centaines ou plusieurs milliers d'arbres.

3.2.3.2 LE GRADIENT BOOSTING

Au lieu d'estimer les B *Weak Learners* en même temps, une alternative consiste à les estimer de façon récursive. Chaque nouvel estimateur est construit de sorte à minimiser l'erreur des précédents estimateurs.

Par exemple, à l'étape $k + 1$ de l'algorithme, k *Weak Learners* ont été calculés. On détermine alors l'erreur de ces modèles :

$$\delta_k = l(Y, g_k(X))$$

$$\text{où } g_k(X) = \sum_{b=1}^k \eta_b * h_b(X)$$

où l est la fonction de perte associée au problème.

Donc le nouveau *Weak Learner* est déterminé tel que :

$$(\hat{\eta}_{k+1}, \hat{h}_{k+1}) = \operatorname{argmin} \sum_{i=1}^n l(\delta_k(i), \eta_k * h_k(x_i))$$

3.2.4 Les hyper-paramètres du modèle

Les **hyper-paramètres** d'un modèle sont des paramètres qui ne sont pas optimisés par l'algorithme mais qui ont un impact significatif sur les résultats s'ils changent. Ils doivent donc être connus et judicieusement choisis par le statisticien.

Nous avons déjà vu des hyper-paramètres dans la sous-partie concernant les arbres CART. Il s'agissait notamment des conditions d'arrêt ou du choix de la mesure de l'homogénéité dans le cas d'un problème de classification.

À ces paramètres s'ajoutent le nombre d'estimateurs B , le sous-échantillonnage des individus et/ou des variables explicatives pour l'entraînement des arbres et le coefficient d'apprentissage (*learning rate* / *shrinkage*).

En général, les hyper-paramètres sont déterminés via des validations croisées dans lesquelles chaque sous-modèle est entraîné avec une valeur différente du paramètre. La valeur choisie de l'hyperparamètre est celle qui minimise la moyenne des erreurs des échantillons de test de la validation croisée.

3.2.5 L'importance des variables dans un modèle de gradient boosting

L'**importance des variables** ou *feature importance* est un score attribué à chaque variable du modèle. Il est utilisé pour comparer l'importance relative des variables explicatives. C'est un score très utile dans la mesure où les modèles ensemblistes sont difficilement interprétables en l'état.

Le **score d'importance** d'une variable dans un arbre de décision correspond au nombre de fois où la variable est sélectionnée dans un *split* et est pondérée par le carré de la performance résultant du *split*.

$$\text{Le score d'importance pour l'arbre } T \text{ et la variable } j \text{ s'écrit : } \hat{I}_j^2(T) = \sum_{t=1}^{J-1} i_t^2 \mathbb{1}_{v_t=j}$$

Où :

- la sommation est réalisation sur tous les nœuds non terminaux, J représente le premier nœud terminal ;
- v_t est la variable associée au nœud t ;
- $\hat{i}_t^2$ correspond à l'amélioration empirique de la performance suite au t -ième *split* ;
- $i^2 = \frac{w_l w_r}{w_l + w_r} * (\bar{y}_l - \bar{y}_r)^2$, avec w_l et w_r les poids des feuilles générées à gauche et à droite et $\hat{y}_l$ et $\hat{y}_r$ sont les moyennes des réponses dans chaque feuilles.

L'importance de la variable sur l'ensemble du modèle correspond à la moyenne des importances pour chaque arbre de décision.

3.2.6 La valeur de Shapley et l'interprétation des boîtes noires

L'inconvénient d'un modèle tel que le gradient boosting est qu'il est difficilement interprétable. En effet, il s'agit d'un modèle agrégeant plusieurs dizaines voire centaines d'arbres de décision.

Dans cette partie est définie la valeur de Shapley qui est une méthode utilisée pour l'interprétation de modèles d'apprentissage. **Le but est de déterminer la contribution de chaque variable explicative sur la prédiction.**

3.2.6.1 LA VALEUR DE SHAPLEY EN THÉORIE DES JEUX

La valeur de Shapley a été initialement introduite en théorie des jeux. Dans un jeu coopératif, elle permet d'attribuer les gains de façon équitable à chaque joueur.

Un jeu coopératif est défini tel que :

- il existe un ensemble N de n joueurs ;
- il existe une fonction $v : 2^N \rightarrow \mathbb{R}$ avec $v(\emptyset) = 0$.

La fonction v est appelée fonction caractéristique du jeu. Elle signifie que pour une coalition $S \subseteq N$ de joueurs $v(S)$ correspond au gain de la coalition. Il est logique que le gain soit nul pour une coalition composée d'aucun joueur ($v(\emptyset) = 0$).

La valeur de Shapley est une façon de distribuer le gain total du jeu entre chacun des joueurs. La valeur de Shapley du i -ième joueur dans un jeu coopératif s'écrit :

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} \times (v(S \cup \{i\}) - v(S))$$

La formule peut être réécrite de la façon suivante :

$$\varphi_i(v) = \frac{1}{N} \sum_{S \subseteq N \setminus \{i\}} \binom{n-1}{|S|}^{-1} \times (v(S \cup \{i\}) - v(S))$$

Et elle peut être interprétée :

$$\varphi_i(v) = \frac{1}{\text{Nombre de joueurs}} \sum_{\text{Coalitions excluant } i} \frac{\text{Contribution marginale de } i \text{ à la coalition}}{\text{Nombre de coalitions excluant } i \text{ de cette taille}}$$

La valeur de Shapley possède de nombreuses propriétés intéressantes qui font d'elle une méthode de répartition des gains équitable :

- la valeur de Shapley est **additive** : $\sum_{i \in N} \varphi_i(v) = v(N)$;
- la valeur de Shapley est **linéaire** : $\varphi_i(a \times v + w) = a \times \varphi_i(v) + \varphi_i(w)$;
- la valeur de Shapley est **symétrique** :
si deux joueurs i et j sont équivalents au sens de $v(S \cup \{i\}) = v(S \cup \{j\})$,
alors $\varphi_i(v) = \varphi_j(v)$;
- la valeur de Shapley est **nulle pour un joueur n'ayant pas d'impact sur le gain**
(i.e. : $v(S \cup \{i\}) = v(S)$).

La valeur de Shapley est la seule distribution de gain qui satisfasse ces quatre propriétés.

3.2.6.2 INTERPRÉTATION DE BOÎTES NOIRES

Plaçons-nous maintenant dans le contexte d'un modèle d'apprentissage. Les variables explicatives sont assimilées aux joueurs du jeu coopératif et la prédiction constitue le gain du jeu. Ainsi, il est possible de calculer la valeur de Shapley pour chacune des variables explicatives. Cette valeur est alors interprétée comme la contribution de la variable sur la prédiction.

La **figure 3.6** montre la contribution de chaque variable explicative. Cette contribution amène à déplacer la prédiction au-delà de sa valeur moyenne.

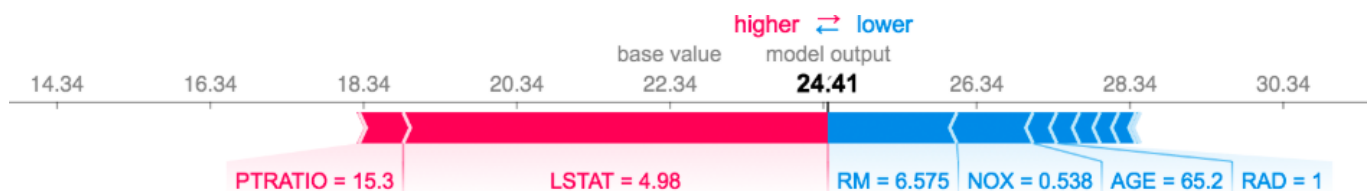


Figure 3.6 – **Contribution des variables explicatives dans un modèle d'apprentissage**

Source : Lundberg Scott, « A unified approach to explain the output of any machine learning model » [en ligne], disponible sur <https://github.com/slundberg/shap> [consulté le 8 septembre 2019]

Lecture :

- La valeur *base value* indique la moyenne du modèle (i.e. : ce que renvoie un modèle sans variables explicatives).
- La valeur *model output* correspond à la valeur prédite par le modèle. Les contributions de chacune des variables ont déplacé la *base value* vers le *model output*.

L'inconvénient de la valeur de Shapley est que le nombre de calculs nécessaires évolue de façon exponentielle en fonction du nombre de variables explicatives. En effet, la valeur de Shapley nécessite de calculer 2^N valeurs, où N est le nombre de variables. Ainsi, dans le cas d'un nombre élevé de variables explicatives, l'estimation de la valeur de Shapley peut être très imprécise.

3.3 Comparaison des modèles

Dans cette partie sont présentés des indicateurs statistiques, de performance et graphiques permettant de comparer différents modèles statistiques entre eux.

3.3.1 Indicateurs statistiques

Deux critères statistiques pour comparer les modèles paramétriques sont présentés dans cette partie.

1. La statistique et le test de Student :

En utilisant les propriétés énoncées dans le paragraphe 3.1.2 sur les estimateurs $\hat{\beta}$ et $\hat{\sigma}$, il est possible de construire la statistique :

$$T = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{\sigma}^2 (X'X)^{-1}_{ii}}} \sim \text{Student}(n - k)$$

Le statistique de Student permet de tester l'hypothèse : $H_0 : \hat{\beta}_i = \beta_i$.

À un niveau de confiance α fixé, la valeur absolue de la statistique est comparée au quantile d'ordre $1 - \frac{\alpha}{2}$ d'une loi de Student à $n - k$ degrés de liberté noté $t_{1-\frac{\alpha}{2}, n-k}$.

Si : $|T| > t_{1-\frac{\alpha}{2}, n-k}$ alors H_0 est rejetée.

En général, les logiciels de statistiques produisent un test de Student testant l'hypothèse : $H_0 : \hat{\beta}_i = 0$ pour chacun des coefficients du modèle.

2. Le critère d'information d'Akaike (AIC) :

Ce critère mesure la vraisemblance d'un modèle tout en pénalisant la complexité de celui-ci.

Il s'écrit :

$$AIC = 2k - 2 \log(L)$$

Où k est le nombre de paramètres du modèle et L est la vraisemblance du modèle par rapport aux données.

L'AIC ne permet pas de connaître la qualité du modèle de manière absolue, il ne permet que de comparer différents modèles et d'en sélectionner un. Le modèle sélectionné est le modèle minimisant l'AIC.

Ce critère permet notamment de limiter le sur-apprentissage, car le modèle sélectionné aura tendance à avoir moins de paramètres et donc à être plus généralisable.

3. La déviance :

Le modèle saturé est défini par :

$$\hat{\mu}_i = y_i ;$$

signifiant que chaque prédiction vaut la valeur observée. Ce n'est pas un modèle très intéressant en pratique, car il ne fait que reproduire les valeurs observées, il possède donc une variance très importante.

Toutefois, il est utilisé comme référence pour comparer plusieurs modèles avec la même densité et la même fonction de lien. Un modèle est considéré comme décrivant correctement les données si la vraisemblance du modèle est proche de celle du modèle saturé.

La déviance correspond à l'écart entre la log-vraisemblance du modèle et celle obtenue avec modèle saturé.

$$D = 2\phi \times [\log(L_{sat.}) - \log(L_{modèle})]$$

Une grande valeur de D traduit un mauvais ajustement du modèle, alors qu'une petite valeur signifie un bon ajustement. De plus, si le modèle décrit bien les données observées, D suit approximativement la loi χ^2_{n-p-1} (où $p+1$ est le nombre de paramètres du modèle).

Ainsi, un modèle est jugé de mauvaise qualité si :

$$D > \chi_{n-p-1;1-\alpha}^2$$

Où $\chi_{n-p-1;1-\alpha}^2$ est le quantile d'ordre $1 - \alpha$ de la loi khi-deux à $n - p - 1$ degré de liberté.

Remarque : Dans le cas où les observations suivent une loi de Bernoulli, la déviance ne permet pas de juger de la qualité du modèle. En effet, dans ce cas, la déviance ne dépend que des valeurs ajustées $\hat{y}_i$ et pas des observations y_i et donc elle ne donne aucun renseignement sur l'ajustement.

Ce mémoire traitant du cas d'observations suivants une telle loi, la déviance ne sera pas utilisée comme indicateur d'ajustement du modèle.

4. Le test du rapport de vraisemblance :

Le test du rapport de vraisemblance permet de comparer des modèles paramétriques emboîtés. Il teste l'hypothèse $H0 : \beta \in \Theta_0$, où Θ_0 représente un sous-espace de l'ensemble des variables explicatives.

La statistique de test s'écrit :

$$\lambda = -2 \log \left(\frac{\mathcal{L}(\hat{\theta}_0)}{\mathcal{L}(\hat{\theta})} \right)$$

Où : $\mathcal{L}(\hat{\theta}_0)$ (resp. $\mathcal{L}(\hat{\theta})$) correspond à la vraisemblance du modèle sous l'hypothèse $H0$.

Sous l'hypothèse $H0$ la statistique λ suit une loi du χ^2 avec un nombre de degrés de liberté égal au nombre de contraintes imposées par l'hypothèse nulle (i.e. : le nombre de variables explicatives dans $\Theta_0 + 1$, que l'on note $k + 1$).

À un niveau de confiance α fixé, la statistique λ est comparée au quantile d'ordre $1 - \alpha$ d'une loi de χ_{k+1}^2 , noté $q_{1-\alpha, k+1}$.

Ainsi $H0$ est rejetée si : $\lambda > q_{1-\alpha, k+1}$.

3.3.2 Indicateurs de performance

La courbe ROC (*Receiver Operating Characteristic*) et l'aire sous celle-ci sont les indicateurs de performance du modèle que nous utiliserons. Il s'agit d'indicateurs classiques dans le cadre des problèmes de classification binaire.

Graphiquement, la courbe représente le taux de vrais positifs (ou sensibilité) en fonction du taux de faux positifs pour tous les seuils possibles.

Chaque probabilité p calculée par le modèle pour chaque individu est comparée à un seuil S . Ainsi si $p > S$ alors l'individu est classé positivement, sinon il est classé négativement. La courbe ROC trace alors les taux de vrais positifs et de faux positifs en faisant varier le seuil S . Le seuil décroît au fur à mesure que la sensibilité augmente.

Pour rappel, le taux de vrais positifs correspond à la fraction des positifs qui sont effectivement détectés. Le taux de faux positifs est la fraction des négatifs qui sont incorrectement détectés.

De façon analogue, sont définis le taux de vrais négatifs (ou spécificité) et le taux de faux négatifs.

Il est possible de remarquer que :

$$\text{Taux de faux positifs} = 1 - \text{Spécificité}$$

Les schémas de la **figure 3.7** ci-dessous résument les concepts qui viennent d'être présentés, ainsi que le classifieur parfait et le classifieur aléatoire de la courbe ROC.

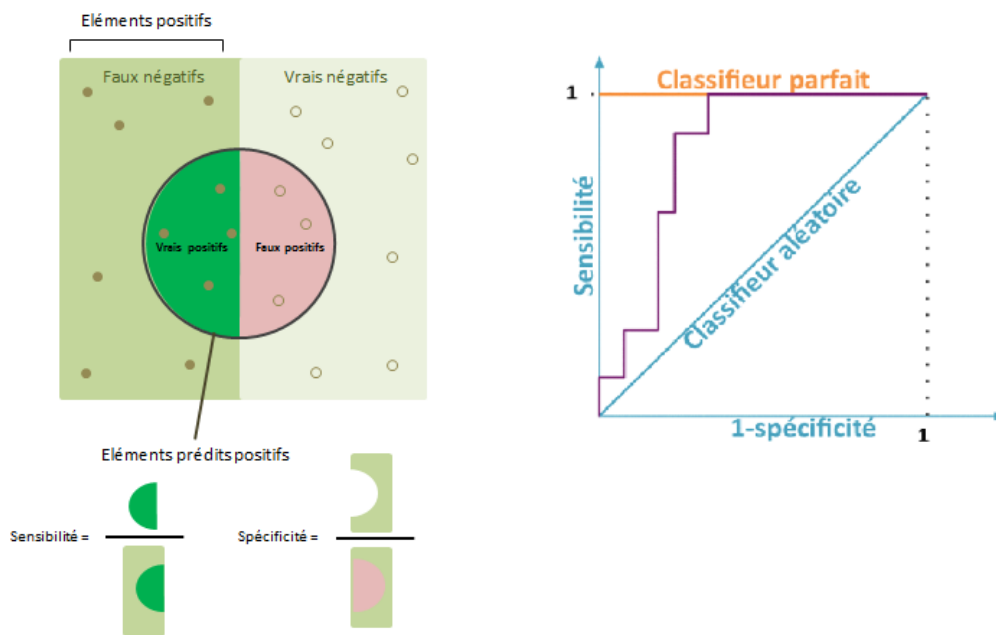


Figure 3.7 – Schéma explicatif du la précision et du rappel (à gauche) et courbe ROC (à droite).

Sources : Walber, 2014 (gauche) et Ch.-A. Azencott, Openclassrooms.com, 2019 (droite).

Au point (0,0) de la courbe ROC, le seuil est trop élevé pour qu'aucune observation ne soit classée positivement. Au point (1,1), le seuil est suffisamment bas pour que toutes les observations soit classées positivement.

Ainsi, plus l'aire sous la courbe est grande, plus le classifieur s'éloigne du classifieur aléatoire et se rapproche du classifieur idéal.

L'aire sous la courbe ROC (AUC : *Area Under Curve*) du classificateur parfait est de 1, et l'AUC sous la courbe du classificateur aléatoire est de 0,5. Par conséquent, l'AUC du modèle permet de mesurer sa proximité aux modèles idéal et aléatoire. En général, une AUC inférieure à 0,7 est considérée comme faible alors qu'une AUC au-delà de 0,9 est considérée comme excellente. Ces considérations sont naturellement amenées à changer en fonction de la difficulté du problème à modéliser.

Avantage : L'avantage de cet indicateur est qu'il s'agit d'une mesure de performance absolue, contrairement à l'AIC. L'AUC est bornée par les valeurs des modèles aléatoires et parfaits.

Inconvénient : L'un des inconvénients de la courbe ROC et de l'aire sous celle-ci est qu'ils sont sensibles aux bases de données déséquilibrées. En effet la courbe ROC prend en compte la capacité à bien prédire les négatifs comme les positifs.

3.3.3 Mesurer la pertinence des prédictions de façon graphique

Dans cette partie, nous proposons une représentation graphique des prédictions. Cette représentation permet d'apprécier le biais et la volatilité des prédictions entre les différents modèles. De plus, elle permet de visualiser la répartition des prédictions et par extension comment le modèle arrive à identifier des classes d'individus à forte ou faible probabilité.

Soient $\hat{P}$ le vecteur des prédictions dans lequel les prédictions sont triées par ordre croissant et P le vecteur des labels.

Nous voulons vérifier si en moyenne le modèle est précis par classe de risque qu'il identifie. Pour ce faire, le vecteur des prédictions est divisé en k classes puis la moyenne des prédictions et des labels par classe est calculée. Ces moyennes sont ensuite tracées sur un graphique pour apprécier visuellement l'écart entre les prédictions et la réalité.

Le nombre de classes est déterminé de sorte à avoir un nombre suffisant d'individus au sein de chacune d'entre elles pour assurer la stabilité statistique des prédictions (une classe de 3 000 individus génère une incertitude inférieure à 1 % sur la moyenne calculée).

Ainsi, la volatilité qui pourrait être constatée entre les moyennes des prédictions et des valeurs réelles s'explique uniquement par la volatilité du modèle et non par le fait que les classes présentent trop peu d'individus.

La **figure 3.12** présente une telle courbe ainsi qu'une interprétation détaillée de celle-ci.

3.4 Application et comparaison des modèles

Dans cette partie, la probabilité que le client se multi-équipe durant l'année est modélisée. Nous cherchons à modéliser la *vraie* probabilité pour chaque individus car dans la suite du mémoire cette probabilité est utilisée pour pondérer la valeur en cas de souscription pour aboutir à la valeur potentielle. Ainsi nous ne tentons pas de rééquilibrer la base de données par des méthodes de sur-échantillonnage, sous-échantillonnage ou de rééquilibrage des poids des individus.

Expliquons la raison avec un exemple dans lequel les poids sont rééquilibrés :

- un individu appartenant à la classe minoritaire a désormais un poids de 4 ;
- un individu appartenant à la classe majoritaire garde quant à lui un poids de 1.

Ici, les individus de la classe minoritaire ont un poids 4 fois plus important par rapport aux autres. Dès lors la proportion d'individus de la classe minoritaire passe de 6 % à 19 % (cela correspond à $\frac{4 \times 6\%}{4 \times 6\% + 94\%}$). La base a bien été rééquilibrée.

Imaginons ensuite qu'un arbre de classification est entraîné sur cette base. Dans cet exemple l'arbre identifie une feuille dans laquelle 2 individus ont le label 1 (classe minoritaire) et 8 individus ont le label 0 (classe majoritaire).

Ainsi dans la mesure où la prédiction correspond à la moyenne des poids, la prédiction dans le cas de la base rééquilibrée vaut :

$$\frac{\text{Poids des labels 1}}{\text{Poids des labels 0} + \text{Poids des labels 1}} = \frac{2 \times 4}{2 \times 4 + 8} = 50 \%$$

Dans le cas de la base non rééquilibrée, la prédiction vaut :

$$\frac{\text{Poids des labels 1}}{\text{Poids des labels 0} + \text{Poids des labels 1}} = \frac{2}{2 + 8} = 20 \%$$

Ainsi lorsque la valeur moyenne en cas de souscription sera pondérée par la valeur prédiction, la valeur potentielle sera grandement surrestimée dans le cas où le rééquilibrage a été effectué.

3.4.1 Le modèle de régression logistique

3.4.1.1 LE MODÈLE SUR LA BASE DE DONNÉES DE 2015-2017

La base de données est séparée en un échantillon d'entraînement (80 %) et un échantillon de validation (20 %). De plus les variables numériques sont centrées-réduites. Les moyennes et variances utilisées pour la réduction sont stockés afin d'appliquer la même transformation aux futures base de données. Centrer-réduire les variables permet d'éviter de donner trop d'importance aux variables avec une variance importante.

La méthode de sélection de variables pour la régression logistique est la pénalisation Lasso.

Avant de pouvoir sélectionner un modèle (et donc sélectionner les variables sous-jacentes à ce modèle), il convient d'entraîner plusieurs modèles en faisant varier la pénalisation λ de la méthode Lasso. Pour chaque valeur de λ plusieurs modèles sont entraînés via une validation croisée à partir de l'échantillon d'entraînement. Le graphique ci-dessous représente en ordonnée l'AUC moyenne sur les échantillons de test en fonction de la pénalisation qui a été mise à l'échelle logarithmique. Sur le second axe des abscisses en haut est indiqué le nombre de variables activées par la pénalisation (le nombre de degrés de liberté du modèle est retrouvé en y ajoutant 1). Par exemple, tout à droite (**graphique 3.8**), aucune variable n'est activée car la pénalisation est trop forte.

Les bornes autour des points rouge indiquent l'incertitude sur la valeur de l'AUC.

Nous remarquons que plus le nombre de variables activées augmente, plus l'impact de celles-ci sur l'AUC est marginal.

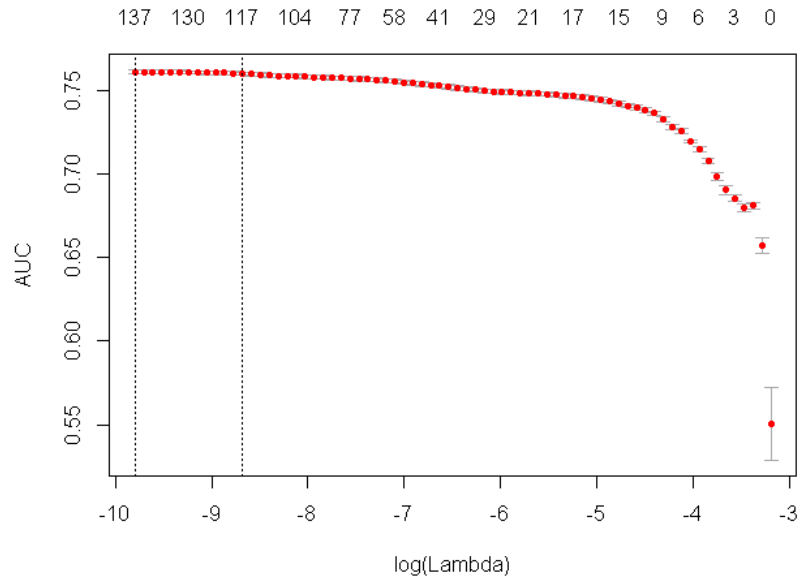


Figure 3.8 – Performance du modèle en fonction de la pénalisation

Nous choisissons de comparer un modèle jugé parcimonieux avec un modèle ayant beaucoup de degrés de liberté (**graphiques 3.9**). Les prédictions du modèle sont calculées sur la base d'entraînement et de validation. La perte d'AUC entre les deux modèles est d'un peu plus de 1 point. De plus, aucun des deux modèles ne sur-apprend. En effet, l'écart entre l'échantillon d'apprentissage et de validation est inférieur à 1 point dans les deux cas.

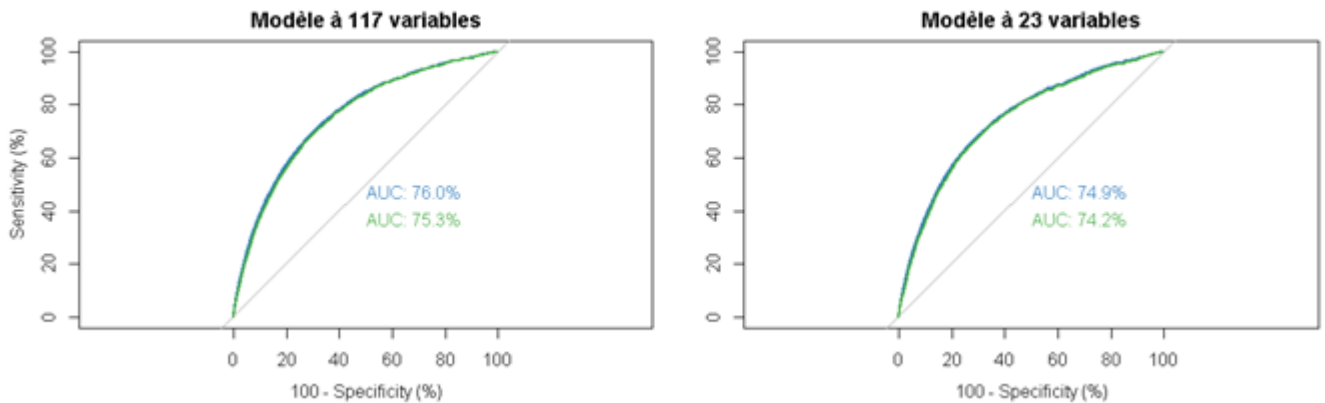


Figure 3.9 – Performances en fonction du nombre de variables sélectionnées

Courbe bleue : train. Courbe verte : test.

Dans la suite, le modèle à 23 variables est utilisé (l'*Intercept* est compté comme une variable) (**tableau 3.10**). La significativité des variables sélectionnées est maintenant testée via le test de Student. Ci-dessous, sont représentées la valeur des coefficients associés aux variables explicatives ainsi que l'erreur et la p -value associée au test de Student. Les variables en rouge sont les variables étant non significatives à plus de 5 %.

Variable	Coefficient	Erreur	Statistique	P value	Significativité
(Intercept)	-3,00	0,04	-71	0%	+++
Ancienneté du client	-0,37	0,01	-45	0%	+++
CSP du client : Artisans, commerçants et chefs d'entreprise	0,05	0,03	1	13%	
Statut occupant : Propriétaire	0,08	0,03	3	1%	++
Indicatrice de renseignement de la banque du client	0,19	0,02	11	0%	+++
Indicatrice de renseignement de la CPS du client	-0,03	0,03	-1	35%	
Indicatrice de renseignement de l'email du client	0,07	0,02	4	0%	+++
Code réseau : Courtiers ou assimilés	-0,43	0,04	-12	0%	+++
Code territoire : Province	0,08	0,02	4	0%	+++
Budget commercial	-0,06	0,01	-7	0%	+++
Produit d'appel : MRH / PNO	0,19	0,04	5	0%	+++
Nombre de contrats détenus	0,10	0,01	12	0%	+++
Indicatrice de multi-équipement	0,17	0,03	5	0%	+++
Possède au moins un contrat Flotte Auto	0,42	0,07	6	0%	+++
Possède au moins un contrat Auto	0,78	0,02	36	0%	+++
Réseau de distribution : Différent du réseau Agent	-0,08	0,04	-2	3%	+
Nombre d'enfants : > 0	0,15	0,03	5	0%	+++
Prime nette totale : > 235	0,05	0,02	3	1%	+
Panier produit : Mono Assurance Vie ou Mono Assurance de Personnes	-0,39	0,03	-14	0%	+++
Délai entre la souscription du 1er et 2ème contrat : > 0 mois	0,26	0,03	8	0%	+++
Délai entre la souscription du 1er et 2ème contrat : Pas de second contrat	-0,08	0,03	-2	2%	+
Délai entre la souscription du dernier et avant-dernier contrat : > 0 mois	0,21	0,03	6	0%	+++
Nombre de contrats détenus : > 3	0,27	0,03	8	0%	+++

Figure 3.10 – Coefficients du modèle à 23 variables

L'hypothèse de nullité des coefficients ne peut être rejetée pour les variables '*CSP : Artisans, commerçants et chefs d'entreprise*' et '*Indicatrice de la CSP*' via la statistique de Student.

Des modèles sont alors entraînés dans lesquels ces deux variables non significatives sont retirées ou maintenues. Quatre modèles sont alors nécessaires. Ensuite, un test du rapport de vraisemblance est réalisé entre le modèle pour lequel les deux variables ont été retirées et chacun des autres modèles. Les résultats des tests sont présentés dans le **tableau 3.11** ci-dessous.

Chaque modèle est codé par deux nombres, le premier fait référence à la variable '*CSP : Artisans, commerçants et chefs d'entreprise*' et le second à la variable '*Indicatrice de la CSP*'. Si un nombre vaut 0 cela signifie que la variable est absente du modèle, sinon cela signifie qu'elle y est présente.

Modèle	Déviance	Gain de déviance par rapport au modèle 000	P value
Modèle 00	127327,9	NA	NA
Modèle 11	127324,3	4	16%
Modèle 10	127327,9	0	78%
Modèle 01	127324,4	4	6%

Figure 3.11 – Test du rapport de vraisemblance

Aucun des modèles contenant au moins l'une des deux variables ne rejette l'hypothèse de non significativité. Ainsi, le modèle choisi est le modèle 00.

L'aire sous la courbe ROC du modèle est de 74,2 % pour la base de test, la suppression des deux variables non significatives n'a pas eu d'effet sur les performances du modèle.

Analyse du modèle de régression logistique : puisque les variables quantitatives ont été centrées et réduites, il est possible de comparer l'impact d'une variable sur les prédictions du modèle en comparant les coefficients associés.

Les variables qui poussent le plus la probabilité prédite vers de fortes valeurs sont :

- le fait de posséder au moins un contrat *Automobile* ;
- le fait de posséder au moins un contrat *Flotte Automobile*.

Les variables qui poussent le plus la probabilité prédite vers de faibles valeurs sont :

- le fait d'appartenir au réseau de distribution '*Coutiers ou assimilés*' qui correspond au réseau de courtiers et assimilés (hors Grands Courtiers) ;
- le fait d'être mono-équipé en produit *Vie* ou d'assurance de *Personnes* ;
- le fait d'avoir une ancienneté forte.

Enfin la courbe qui a été présentée dans la **partie 3.3.3** est tracée pour la base test (**graphique 3.12**).

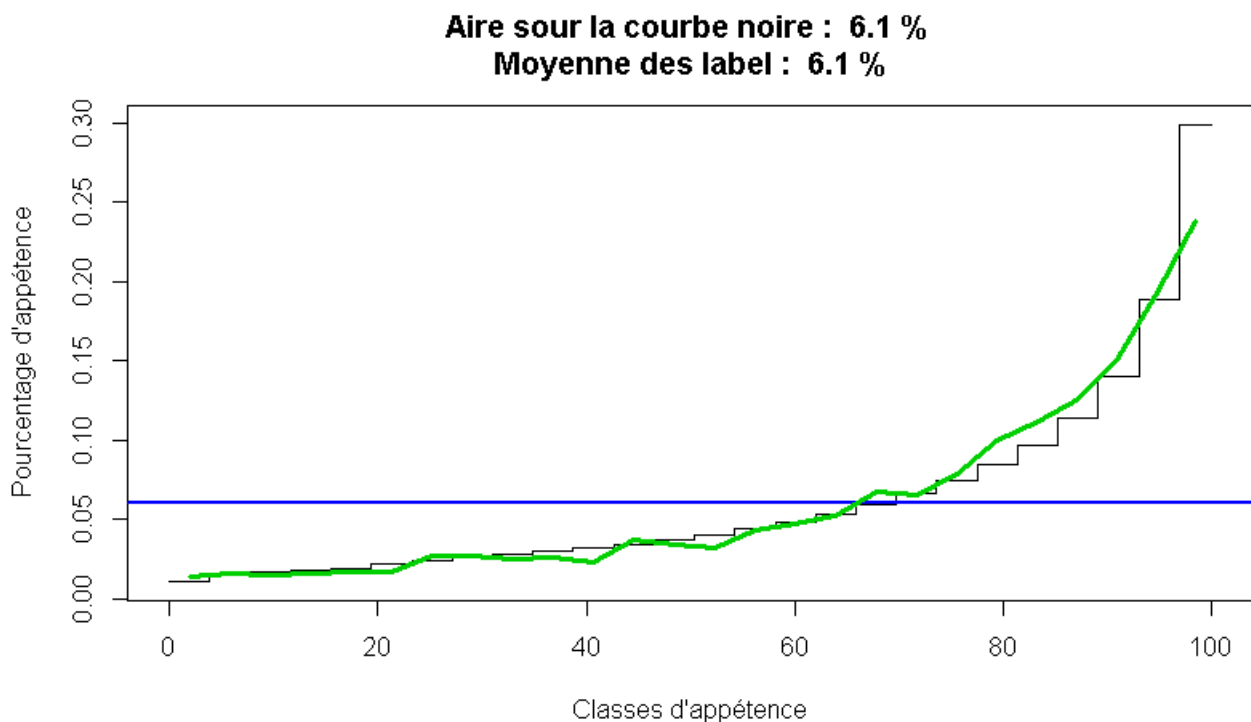


Figure 3.12 – **Comparaison des prédictions et des valeurs réelles**

Les prédictions ont été regroupées en 26 classes de probabilités homogènes. Chaque classe contient 3000 observations sauf la dernière qui en contient 2416. La courbe noire correspond

à la probabilité moyenne au sein de chaque regroupement : cette courbe est croissante car les prédictions ont été rangées dans l'ordre croissant. La courbe verte correspond à la moyenne des labels pour chaque regroupement. La droite verticale bleue correspond au modèle sans variable explicative, il s'agit du modèle prédisant la moyenne de la variable réponse pour tous les individus de la base de données. L'axe des abscisses représente la répartition des individus de la base de données par classe. Par exemple nous pouvons lire que 20 % des individus ont une probabilité prédite inférieure à 2,5 %.

Ainsi, en comparant les prédictions moyennes par classes au modèle sans variables (courbe bleue) nous constatons que 30 % des individus ont une probabilité de se multi-équiper supérieure à la moyenne. De plus, les probabilités pour ces individus sont légèrement sous-estimées car la courbe des taux réels (verte) est légèrement supérieure à la courbe des taux modélisés (noire). En outre, la classe présentant le potentiel le plus élevé a un pourcentage d'appétence de 30 %.

La courbe verte permet de mesurer l'ajustement du modèle : si celle-ci est proche de la courbe noire alors la biais est faible. De plus, la volatilité de la courbe verte traduit une instabilité des prédictions.

Dès lors, le modèle ne crée pas de biais car les deux courbes sont très proches. De plus, l'aire sous chacune des courbes vaut 6,1 %, cette aire correspond à la moyenne des prédictions (respectivement labels), il n'y a donc pas de biais.

Enfin, la courbe verte est peu volatile : cela signifie que la variance de la prédiction moyenne par classe est peu élevée.

3.4.1.2 LE MODÈLE SUR LA BASE DE DONNÉES DE 2017-2018

Le modèle qui vient d'être entraîné est utilisé sur la base de données de 2017-2018. Cela a pour but de vérifier que le modèle reste stable dans le temps. Pour rappel la base de données de 2017-2018 présente un taux moyen de multi-équipement de 4,9 %.

La **figure 3.13** représente la courbe ROC sur cette base de données.

Nous constatons une amélioration de l'aire sous la courbe par rapport aux résultats sur les bases d'entraînement et de test. Toutefois, cela ne doit pas être interprété comme une amélioration des résultats. En effet, nous avons expliqué dans la **partie 3.3.2** que la courbe ROC est un indicateur sensible aux bases de données déséquilibrées. Or la base de 2017-2018 est plus déséquilibrée que la base d'entraînement, c'est pourquoi l'aire sous la courbe ROC a augmenté.

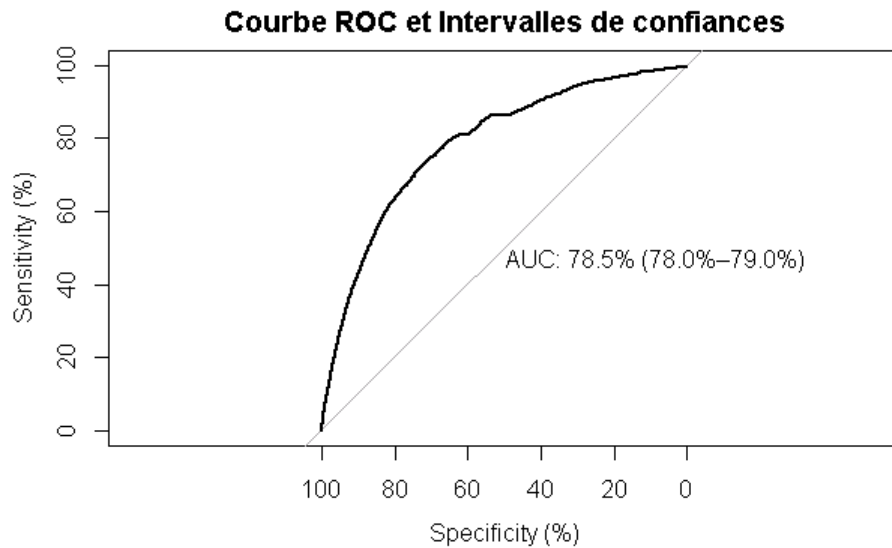


Figure 3.13 – Comparaison des prédictions et des valeurs réelles

L'observation de la **figure 3.14** permet de vérifier l'ajustement du modèle par rapport à la base de données.

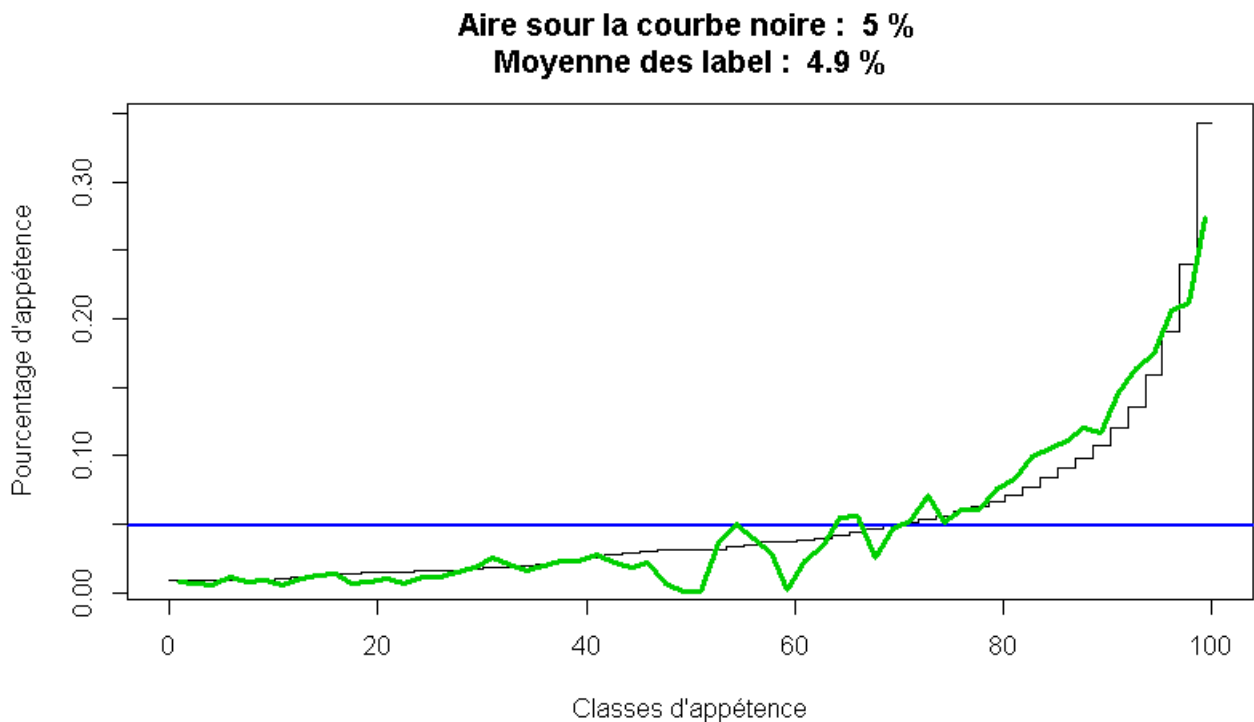


Figure 3.14 – Comparaison des prédictions et des valeurs réelles

Nous constatons que le biais est faible. Toutefois, le modèle est volatile : nous observons 3 cassures sur la courbe verte. Ces cassures se produisent au milieu de la distribution entre le percentile 50 % et le percentile 70 %. Cela concerne donc des individus dont la probabilité est proche de la moyenne. Ainsi, nous voyons que le modèle reste robuste pour les classes d'individus à faible et forte probabilités.

3.4.2 Le modèle de gradient boosting

3.4.2.1 LE MODÈLE SUR LA BASE DE DONNÉES DE 2015-2017

Pour ce modèle, une validation croisée est réalisée sur toute la base de données pour déterminer les hyperparamètres optimaux, il s'agit des hyperparamètres pour lesquels les performances sont les meilleures. La validation croisée permet d'éviter de biaiser la valeur de ces paramètres.

La validation croisée consiste à séparer la base de données en K bases (ici K égal 6). K modèles sont ensuite entraînés, la base d'entraînement et de validation changeant à chaque fois de sorte à balayer les K bases. Le score de test correspond à la moyenne des performances des modèles sur leur base de test respective.

Les deux hyper-paramètres les plus importants sont la profondeur de l'arbre et le nombre d'arbres. Il s'agit des paramètres que nous cherchons à optimiser. Pour ce faire, les performances des modèles, déterminées par validation croisée, en fonction de la profondeur et du nombre d'arbres, sont tracées. Le couple optimal (*profondeur, nombre d'arbres*) correspond au couple maximisant la performance.

Le **graphique 3.15** ci-dessous représente les performances des modèles au sens de l'aire sous la courbe ROC (AUC). Le couple optimal de paramètres correspond à une profondeur de 6 et un nombre d'arbres de 280. L'aire sous la courbe ROC du modèle associé à ces paramètres vaut **78 %**.

De plus, nous constatons que les modèles avec une profondeur supérieure à 3 sur-apprennent pour un nombre d'arbres grand : en effet, la performance chute de façon plus ou moins brutale en fonction de la profondeur.

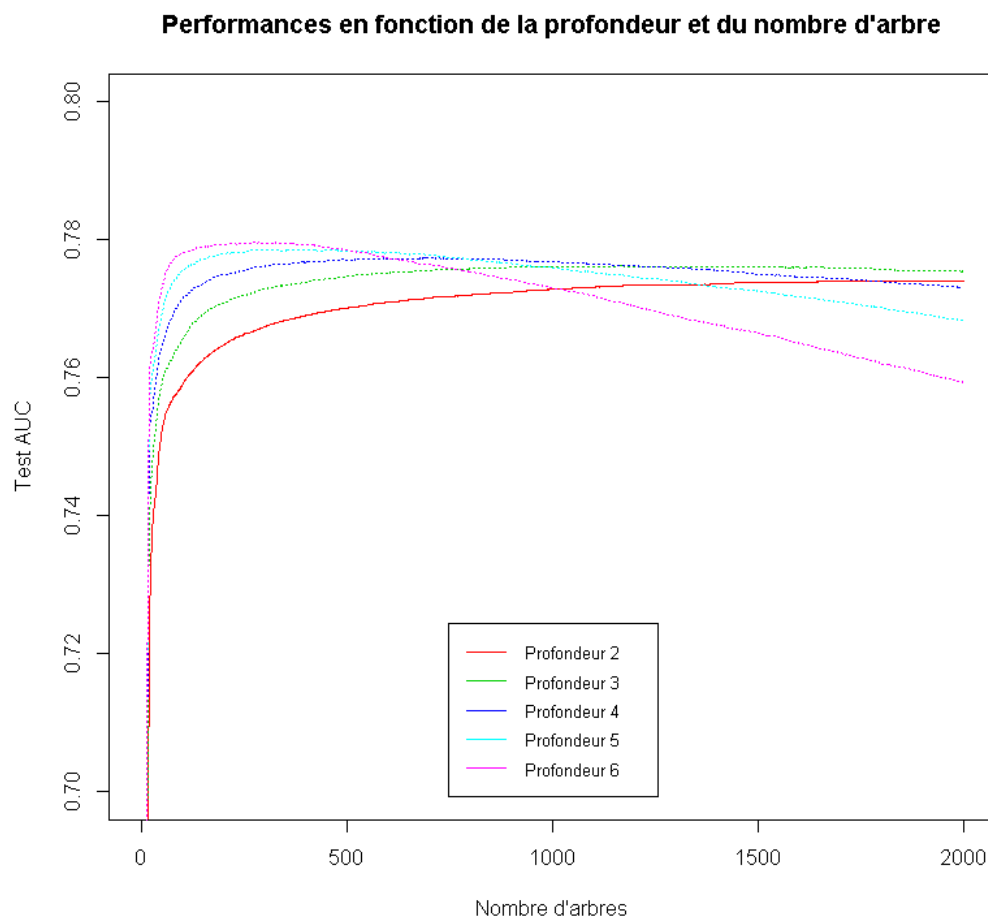


Figure 3.15 – Moyenne des performances sur l'échantillon de test en fonction de la profondeur des arbres et du nombre d'arbres lors de la validation croisée

L'importance des variables est calculée et le **graphique 3.16** représente les quinze variables les plus importantes.

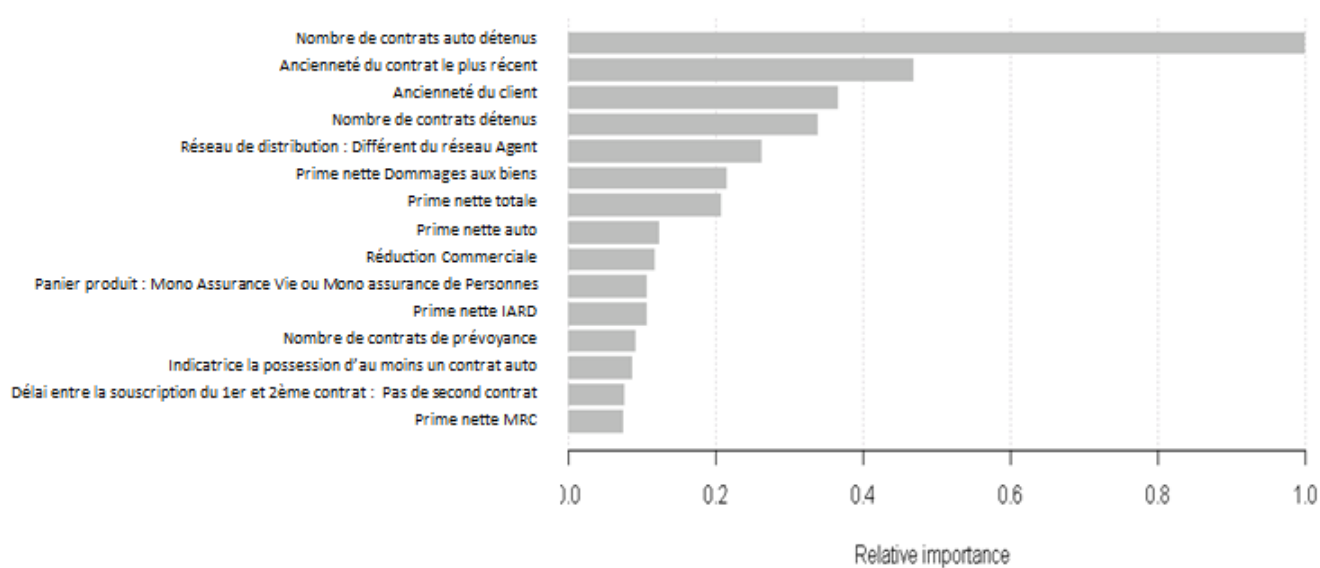


Figure 3.16 – Importance des variables pour les hyperparamètres choisis

Les variables qui ont l'impact le plus important sont :

1. *le nombre de contrats Automobile* ;
2. *l'ancienneté du contrat le plus récent* ;
3. *l'ancienneté du client* ;
4. *le nombre de contrats détenus*.

La variable *nombre de contrats Automobile* est deux fois plus importante que la variable *ancienneté du contrat le plus récent* qui est la seconde variable la plus importante du modèle. De plus l'importance des variables chute rapidement.

Parmi les quatre variables les plus importantes, deux sont aussi sélectionnées par la pénalisation Lasso dans le modèle précédent (*nombre de contrats détenus* et *ancienneté du client*).

Remarques :

- L'importance des variables n'apporte pas d'information au sujet du lien entre la variable et la prédiction (i.e. si la variable discrimine des individus ayant un fort ou faible potentiel).
- Certaines variables liées aux différents montants de prime ont de l'importance dans le modèle (*Prime DAB*, *Prime totale*, *Prime Automobile* ou *Prime IARD*). Cela s'explique par le fait que le modèle de gradient boosting est capable de prendre en compte des relations non linéaires contrairement au modèle de régression logistique. C'est pourquoi la pénalisation Lasso n'a pas sélectionnée de telles variables.

Enfin, nous observons l'ajustement du modèle (**figure 3.17**) :

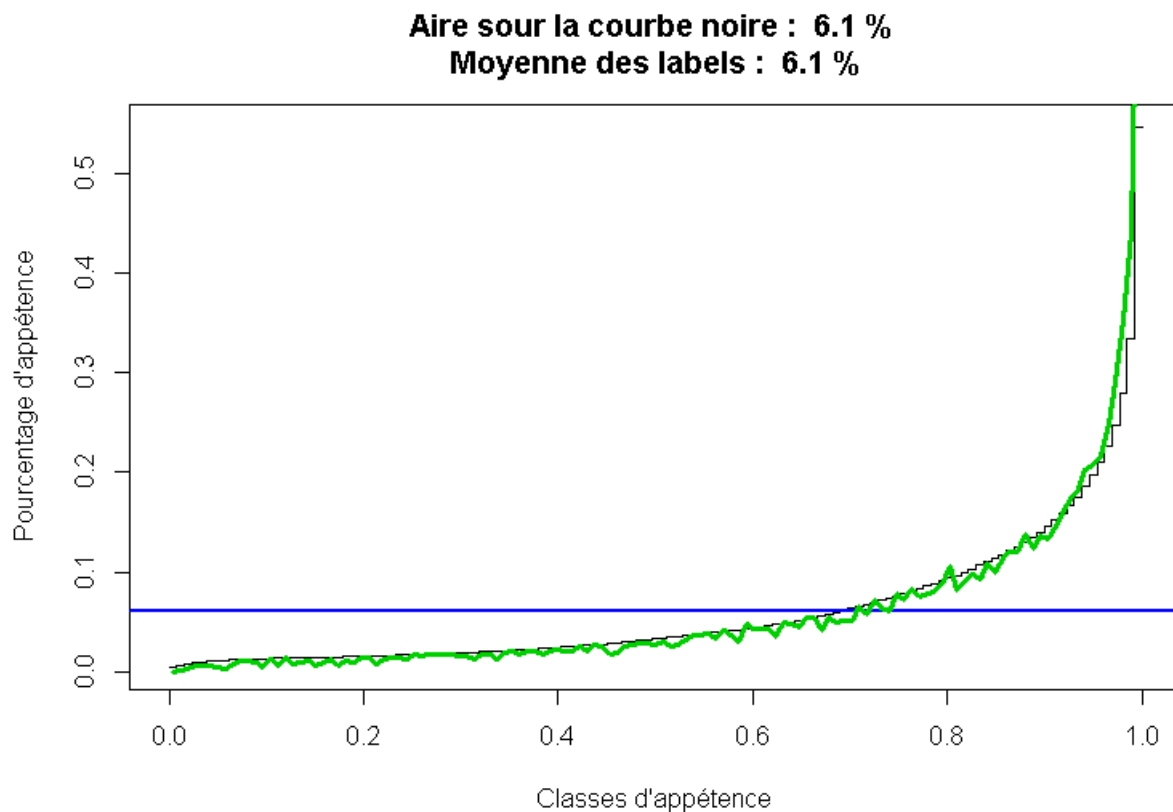


Figure 3.17 – Comparaison des prédictions et des valeurs réelles

Cette courbe n'est pas informative concernant la volatilité du modèle car elle a été réalisée sur la base d'entraînement dans la mesure où il n'y a pas eu de base de test. Cependant nous constatons que la courbe présente une croissante exponentielle plus prononcée que dans le cas du modèle logistique (cf **figure 3.12**). De plus, la probabilité prédite est plus élevée pour les classes d'appétence forte.

3.4.2.2 LE MODÈLE SUR LA BASE DE DONNÉES DE 2017-2018

La courbe ROC et l'aire sous celle-ci sur la base de données de 2017-2018 sont représentées sur le **graphique 3.18** ci-dessous. L'AUC vaut 81 %.

Il a été expliqué, dans l'analyse du **graphique 3.13** que l'augmentation de l'aire sous la courbe s'explique par le fait que la base de données de 2017-2018 est plus déséquilibrée.

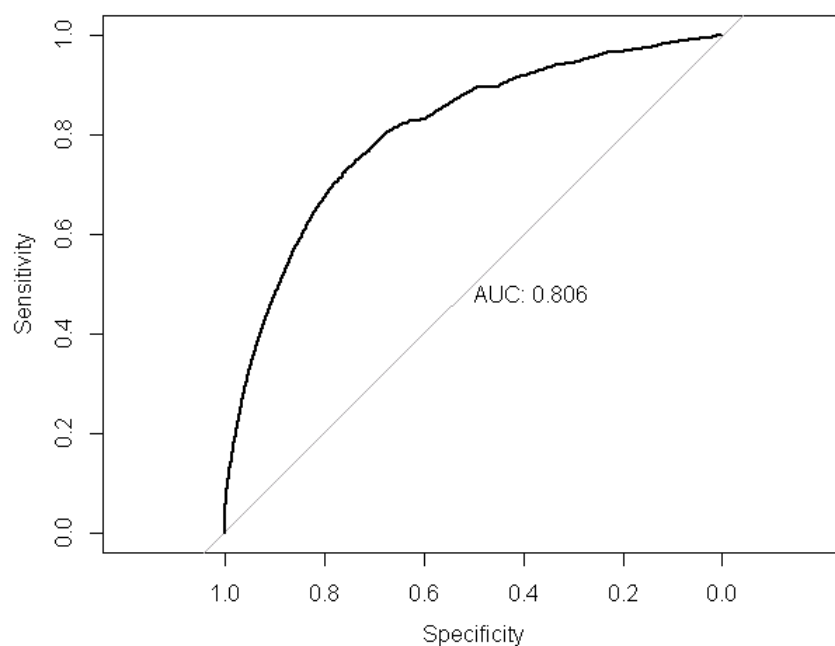


Figure 3.18 – Courbe ROC sur les données de 2017-2018

Observons maintenant l'ajustement du modèle via la **figure 3.19** :

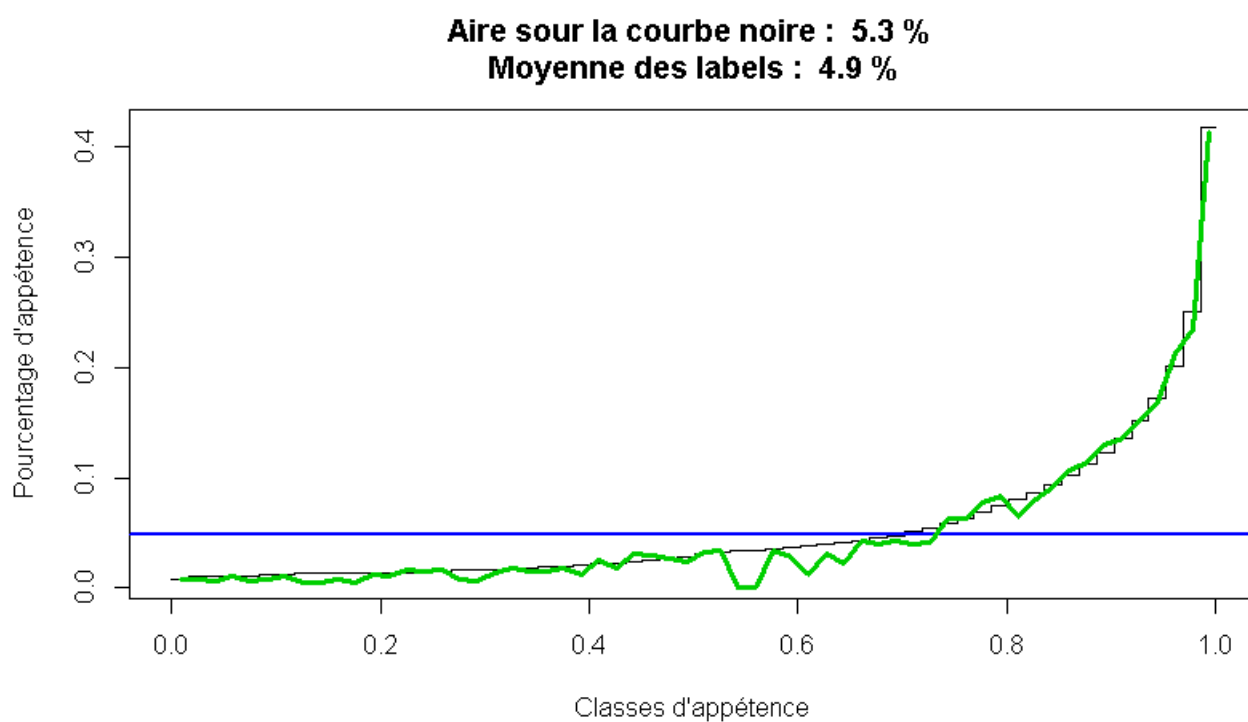


Figure 3.19 – Comparaison des prédictions et des valeurs réelles

Remarque : La **figure 3.19** présente autant de classes que la **figure 3.14**.

Le modèle de gradient boosting présente un biais plus important que le modèle de régression logistique. En effet, la moyenne des prédictions (aire sous la courbe noire) est de 5,3 % alors que la moyenne des labels est de 4,9 %. Toutefois, la courbe des taux réels (courbe verte) est beaucoup moins volatile. Elle ne présente que deux légères cassures entre le percentile 50 % et le percentile 65 %. Comme dans le cas du modèle de régression logistique, le modèle de gradient boosting présente des instabilités concernant les individus dont les prédictions sont proches de la moyenne. Outre ces individus, l'ajustement est très bon sur les individus à forte et faible probabilités.

3.4.3 Conclusion

3.4.3.1 PERFORMANCES DES ANCIENS SCORES DU MODÈLE DE VALEUR POTENTIELLE

Pour rappel, le modèle actuel de valeur potentielle utilisé par le service Valeur Client de Generali est composé de sept scores d'appétence au multi-équipement (*partie 1.1.3*). Les performances du modèle actuel sont calculées et résumées dans le **tableau 3.20** ci-dessous.

Modèle calculant la :	Score de performance : Aire sous la courbe ROC
probabilité de multi-équipement vers les contrats Multirisques habitation (MRH) pour un client détenteur de n'importe quel contrat (hors réseau salarié)	82%
probabilité de multi-équipement vers les contrats MRH pour un client détenteur de n'importe quel contrat (réseau salarié)	74%
probabilité de multi-équipement vers les contrats Multirisques commerce (MRC) pour un client détenteur d'au moins un contrat Automobile (hors réseau salarié)	78%
probabilité de multi-équipement vers les contrats Epargne Retraite pour un client détenteur d'au moins un contrat Automobile et / ou MRH (hors réseau salarié)	75%
probabilité de multi-équipement vers les contrats Automobile pour un client détenteur d'au moins un contrat MRC (hors réseau salarié)	70%
probabilité de multi-équipement vers les contrats Automobile pour un client détenteur d'au moins un contrat MRH (hors réseau salarié)	87%
probabilité de multi-équipement vers les contrats Epargne Retraite pour un client détenteur de n'importe quel contrat (réseau salarié).	71%

Figure 3.20 – Aire sous la courbe ROC des scores de multi-équipement actuels

Les deux modèles développés dans ce mémoire ont une aire sous la courbe ROC supérieure à 5 des 7 modèles actuels. De plus, ils sont plus exhaustifs dans le sens où il considère le multi-équipement vers n'importe quel produit.

3.4.3.2 COMPARAISON DES MODÈLES

Le modèle de gradient boosting présente une aire sous la courbe ROC sur les données de 2017-2018 plus élevée que le modèle de régression logistique. De plus, les graphiques comparant les prédictions et les valeurs réelles montrent que le modèle de gradient boosting est plus stable.

Ainsi, c'est ce modèle qui est gardé pour la suite du mémoire.

Chapitre 4

Modélisation de la rentabilité potentielle du contrat

Le but de ce chapitre est de modéliser la rentabilité des contrats futurs souscrits par le client. La distribution de cette valeur sera calculée pour chaque client. La distribution apporte plus d'information qu'une simple valeur moyenne et elle permet de prendre en compte l'hétérogénéité de valeur en fonction des contrats souscrits.

La première partie a vocation à proposer une formalisation du modèle développé (**partie 4.1**). La seconde partie sera consacrée d'une part à la détermination des produits pour lesquels la valeur sera modélisée de façon plus fine (**partie 4.2.1**) et d'autre part au calcul des probabilités d'équipement vers les produits (**partie 4.2.2**). Ensuite, la méthode pour la modélisation de la distribution sera présentée puis appliquée (**partie 4.3**). Cette méthode se décompose en deux parties : tout d'abord des classes d'individus sont identifiées via des arbres de régression, puis des lois de probabilité sont ajustées sur chaque classe.

4.1 De la modélisation de la valeur moyenne à la modélisation de la distribution

Dans cette partie nous détaillons comment nous avons abouti au choix de modéliser la distribution de la valeur en cas de souscription.

Initialement nous comptons modéliser la valeur moyenne en cas de souscription :

Valeur moyenne en cas de souscription du client ayant pour caractéristiques $x = \sum_{i \in \text{Produits}} p_i^x \times \bar{V}_i^x$

Où :

- p_i^x est la probabilité que le client se multi-équipe du produit i sachant qu'il se multi-équipe (pour tout $x : \sum_i p_i^x = 1$) ;
- $\bar{V}_i^x$ est la valeur moyenne du produit i pour un client ayant les caractéristiques x .

Toutefois il a été constaté que la dispersion autour de la moyenne des contrats $\bar{V}_i^x$ est très forte, cela résulte du fait que les variables utilisées sont peu explicatives.

Le choix a alors été fait de modéliser la distribution afin de quantifier cette dispersion autour de la moyenne. La distribution n'est pas modélisée pour tous les produits mais uniquement pour les produits les plus importants en termes de souscription (cela sera l'objet de la **partie 4.2.1**).

Ainsi nous aboutissons à la densité de distribution de valeur pour un client :

$$\text{Densité de valeur en cas de souscription} = \sum_{i \in \text{Produits}} p_i^x \times \hat{f}_i^x$$

Où :

- p_i^x est la même probabilité que précédemment ;
- $\hat{f}_i^x$ est la densité modélisée au sein de la classe.

4.2 Choix des produits dont la densité sera modélisée et probabilité d'équipement

4.2.1 Détermination des produits dont la densité de la valeur est modélisée

La modélisation de la densité se décompose en deux étapes :

1. Les individus sont découpés en groupes au sein desquels la valeur du contrat est homogène ;
2. La densité est modélisée au sein de chaque groupe.

Ainsi, pour que les groupes puissent être pertinents il faut que le nombre d'individus au sein de chacun d'eux soit conséquent. C'est pourquoi les produits choisis seront les produits vers lesquels les clients se multi-équipent le plus. D'autre part, il doit s'agir de produits que l'équipe *Valeur Client* valorise. En effet, nous constatons que certains produits qui ont un poids important dans le multi-équipement ne sont pas valorisés ou le sont partiellement : notamment en fonction de la génération, des garanties ou du réseau de distribution. Cela s'explique par le fait que l'équipe *Valeur Client* est encore récente, le processus de valorisation n'est encore totalement déployé sur l'ensemble du portefeuille de Generali.

Pour le choix de ces produits, nous nous concentrons uniquement sur le multi-équipement réalisé en 2018 par les clients de moins d'un an d'ancienneté. Cela représente un total de 28 669 contrats vers lesquels les clients se multi-équipent. Sur la **figure 4.1** ci-dessous, le graphique de gauche représente la répartition des produits vers lesquels le client se multi-équipe au cours de l'année 2018 et le tableau de droite indique le pourcentage de contrats qui sont valorisés pour chacun de ces produits.

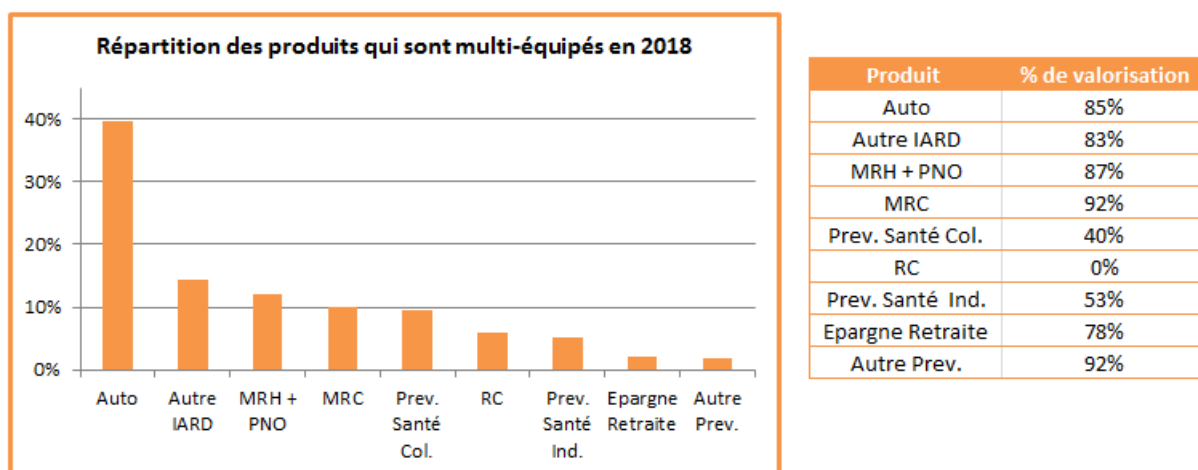


Figure 4.1 – Répartition des produits souscrits par les clients multi-équipés et leur valorisation

D'après le graphique, les produits qui sont les plus multi-équipés sont, en large majorité, les produits *Automobile*, ainsi que les produits des catégories *Autre IARD*, *MRH & PNO*, *MRC* et *Prévoyance santé collective*. Ces quatre derniers produits sont équipés dans des proportions assez similaires, autour de 10 %.

La multiplication entre le pourcentage de valorisation et le pourcentage d'équipement (pourcentages sur le graphique) nous informe de la part de valorisation parmi la répartition des produits équipés. Ces chiffres sont résumés dans le **tableau 4.2** ci-dessous.

Produit		% Répartition des produits qui sont multi-équipés et valorisés
Produits valorisés	Auto	33%
	Autre IARD	12%
	MRH + PNO	10%
	MRC	9%
	Prev. Santé Col.	4%
	RC	0%
	Prev. Santé Ind.	3%
	Epargne Retraite	2%
	Autre Prev.	2%
	Produits non valorisés	25%

Figure 4.2 – Répartition des produits souscrits par les clients multi-équipés et valorisation - Suite

Ainsi, nous constatons que 25 % des produits vers lesquels les clients se multi-équipent ne sont pas valorisés. Le choix initial des produits pour lesquels la densité de valeur est modélisée se porte alors sur :

- *Automobile*
- *MRH & PNO*

- *MRC*

La catégorie *Autre IARD* n'est pas choisie car elle regroupe près de 11 produits qui n'ont pas la même répartition de valeur.

Remarques importantes :

- **Concernant les produits *MRC* :**

La tarification actuelle de ce produit n'est pas en adéquation avec les tarifs de la concurrence. La compagnie a fait le choix d'aligner ses tarifs avec ceux du marché notamment en effectuant des offres commerciales lorsque le client présente un devis de la concurrence avec un tarif inférieur. Ainsi, près de 60 % des produits soucrits en *MRC* se voient attribuer des offres commerciales. Cela a pour effet que les primes perçues ne correspondent pas à la sinistralité modélisée, entraînant une forte dégradation de la valeur de ces produits (plus d'un quart des contrats ont une valeur négative).

Le choix est alors fait de ne pas valoriser ces produits *MRC* dans la suite, car leur valeur actuellement calculée ne représente pas leur rentabilité réelle.

- **Concernant les produits dont la valeur n'est pas modélisée**, certains de ces produits sont regroupés en catégorie de produits. Ces catégories sont :
 - *Autre IARD* qui rassemble les produits *Autre IARD* et la *RC* ;
 - L'*Epargne Retraite* qui est la même catégorie de produits que lors de sa définition dans le chapitre 2 ;
 - La catégorie *Assurance de Personnes*, qui rassemble tous les produits d'assurance de personnes : *Prévoyance* et *Santé* ;
 - La *MRC*, qui forme sa propre catégorie de produits.

De plus, pour ces mêmes produits, il a été choisi d'attribuer la valeur médiane par catégorie de produits.

Le choix de la médiane plutôt que la moyenne s'explique par le fait que cet indicateur est moins sensible aux valeurs extrêmes. En effet, il existe quelques contrats avec des valeurs très importantes qui ne sont pas représentatives du reste de la distribution, mais qui malgré tout tirent la moyenne vers le haut.

De même, le choix de la médiane permet d'avoir une valeur plus prudente dans le cas de catégories qui regroupent plusieurs produits pour lesquels les distributions de valeur sont très différentes. C'est notamment le cas de la catégorie *Autre IARD*, qui regroupe 12 produits, dont la valeur médiane vaut 948 euros contre une valeur moyenne de 3 901 euros. Ce phénomène est retrouvé dans les catégories *Prévoyance* et *Epargne Retraite*.

Ainsi, les valeurs médianes par catégorie sont :

- *Autre IARD* : 948 euros ;
- *MRC* : 87 euros ;
- *Assurance de Personnes* : 786 euros ;
- *Epargne Retraite* : 342 euros.

- **Concernant les produits *Automobile*** : à partir de maintenant, sera introduit le produit *Engins Spéciaux*. En effet, lors de la valorisation, l'*Automobile* et les *Engins Spéciaux* sont distingués. Cette distinction sera respectée dans notre modélisation de la distribution de la valeur. La désignation "*Auto. hors ES*" désignera les produits *Automobile* hors *Engins Spéciaux*.

Conclusion : Les produits pour lesquels la densité de la valeur est modélisée sont :

- "*Auto. hors ES*";
- "*Engins Spéciaux*";
- "*MRH & PNO*".

4.2.2 Détermination du panier produit en cas de souscription

Il a été montré dans le chapitre 2 que le produit d'appel du client donne des informations sur les produits qu'il va souscrire par la suite. De la même façon, il est possible de déduire que la composition du panier produit influence le choix des produits qui sont ensuite souscrits.

Dans cette partie, la composition moyenne du panier produit est déterminée en fonction de certaines variables explicatives, c'est-à-dire que l'on veut déterminer pour un client donné quelle est la probabilité qu'il souscrive tel produit sachant qu'il va s'équiper. Comme dans la partie précédente, nous n'analysons que le multi-équipement réalisé en 2018.

Les variables explicatives choisies sont :

- **La nature du client** : la nature du client est modifiée pour n'être composée que de deux modalités, professionnel ou non-professionnel. La modalité "non-professionnel" est, en majorité, composée de clients ayant pour nature "Entreprise" ou "Petite Entreprise".
- **Le panier produit du client** (avant la souscription du nouveau produit). Cette variable est séparée en sept modalités, qui sont :
 1. le panier produit est uniquement composé d'assurances *Vie*;
 2. le panier produit est uniquement composé d'assurance de *Personnes*;
 3. le panier produit a une composition mixte, c'est-à-dire que le client possède des produits appartenant à au moins deux familles (*IARD*, *Vie* ou de *Personnes*);
 4. le panier produit est uniquement composé d'assurances *IARD* et le client ne possède ni produit *Automobile*, ni produit "*Autre IARD*";
 5. le panier produit est uniquement composé d'assurances *IARD* et le client possède au moins un produit *Automobile*, mais pas de produit "*Autre IARD*";
 6. le panier produit est uniquement composé d'assurances *IARD* et le client possède au moins un produit "*Autre IARD*" mais pas de produit *Automobile*;
 7. le panier produit est uniquement composé d'assurances *IARD* et le client possède au moins un produit *Automobile* et au moins un produit "*Autre IARD*".

Remarque : La catégorie "*Autre IARD*" est la même que celle définie dans la partie précédente.

Le croisement des deux variables "*Nature du client*" et "*Panier produit du client*" permet d'aboutir à 14 classes. La probabilité des produits équipés au sein de chaque classe est calculée. Un nombre restreint de classes, seulement 14, a été choisi de façon à ce que les moyennes calculées puissent présenter une robustesse statistique (voir en annexe A).

Le **tableau 4.3** ci-dessous présente la probabilité de la souscription des produits en fonction de la classe du client.

Remarque : Dans les classes pour lesquelles le client ne possède que des produits *IARD*, le premier chiffre fait référence aux contrats *Automobile* et le second chiffre fait référence aux contrats "*Autre IARD*". Lorsque ce chiffre vaut 1, cela signifie que le client possède au moins l'un des produits, sinon le chiffre vaut 0.

Classe du client	Probabilité de souscrire le produit :						
	Auto hors E.G.	Autre IARD	Engins Spéciaux	Epargne Retraite	MRC	MRH + PNO	Prévoyance
Non PRO - IARD 00	24%	26%	4%	1%	23%	19%	4%
Non PRO - IARD 01	12%	57%	9%	1%	13%	8%	1%
Non PRO - IARD 10	53%	18%	9%	1%	14%	4%	2%
Non PRO - IARD 11	38%	27%	19%	1%	9%	3%	2%
Non PRO - Mixte	39%	23%	9%	4%	13%	6%	6%
Non PRO - Assurance de Pers.	12%	14%	0%	10%	14%	2%	47%
Non PRO - Assurance Vie	4%	5%	0%	73%	7%	0%	12%
PRO - IARD 00	43%	13%	3%	1%	8%	18%	15%
PRO - IARD 01	29%	25%	15%	0%	4%	17%	10%
PRO - IARD 10	52%	9%	5%	0%	3%	23%	7%
PRO - IARD 11	43%	13%	12%	0%	4%	22%	6%
PRO - Mixte	36%	10%	4%	3%	3%	23%	21%
PRO - Assurance de Pers.	13%	5%	1%	5%	2%	11%	63%
PRO - Assurance Vie	10%	1%	0%	27%	1%	2%	60%

Figure 4.3 – Probabilité de souscription des produits équipés en fonction de la classe du client

Nous constatons que la distinction entre les clients professionnels et les clients non-professionnels est justifiée. En effet, les clients professionnels ont une appétence beaucoup plus forte pour les produits *MRH* & *PNO*. Les clients non-professionnels ont, quant à eux, une appétence plus forte pour les produits *MRC*.

4.3 Modélisation de la distribution de valeur par client

La modélisation de la distribution se décompose en deux étapes :

1. Les individus sont découpés en groupe au sein desquels la valeur du contrat est homogène ;
2. La densité est modélisée au sein de chaque groupe.

4.3.1 Création de classes de valeurs via des arbres de régression

Pour rappel, les produits dont la densité de valeur est modélisée sont l' "*Automobile hors E.S.*", les *Engins Spéciaux* et la *MRH* & *PNO*.

Ainsi, trois modèles d'arbres de régression vont être réalisés, le premier pour les produits *Automobile hors E.S.*, le second pour les produits *Engins Spéciaux* et le dernier pour les produits *MRH & PNO*.

La méthode de construction des arbres CART a été présentée dans la **partie 3.2.2**.

Pour rappel, les bases utilisées pour la construction des arbres de régression recensent toutes les affaires nouvelles réalisées en 2018. De plus, la valeur des contrats a été recalculée sans prendre en compte les potentielles offres commerciales (cf. **partie 2.4**).

Les arbres de régression présentent tous les mêmes paramètres qui sont :

1. la profondeur maximale est de 5 ;
 2. le nombre d'individus par feuille doit être supérieur à 5 % de la base de données.
- Ce paramètre vise à réduire le sur-apprentissage.

4.3.1.1 ARBRE DE RÉGRESSION POUR LE PRODUIT *Automobile hors E.S.*

L'**histogramme 4.4**, ci-dessous, représente la répartition de la valeur des contrats *Automobile* hors offres commerciales. La valeur de ces contrats est concentrée autour de 330 euros, mais présente aussi une queue de distribution épaisse. Il sera observé par la suite que la densité au sein des feuilles de l'arbre de régression présente une distribution asymétrique.

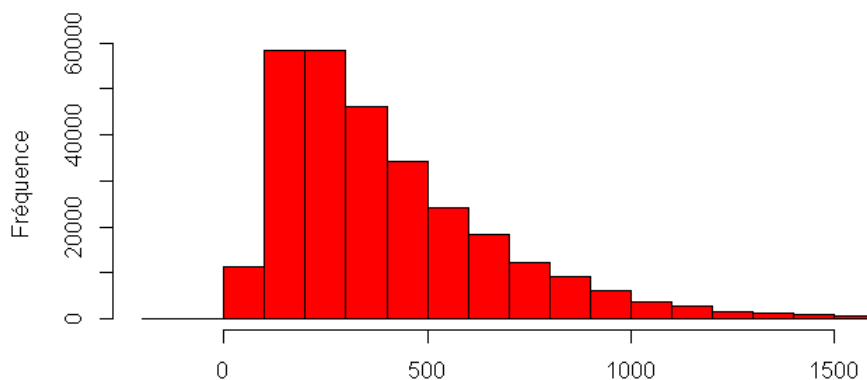


Figure 4.4 – Répartition de la valeur hors offre commerciale des contrats *Automobile hors E.S.*

Ci-dessous (**figure 4.6**) est représenté l'arbre de régression sur la valeur des contrats.

Expliquons la lecture d'un tel arbre à travers l'exemple du premier nœud (**figure 4.5**) :

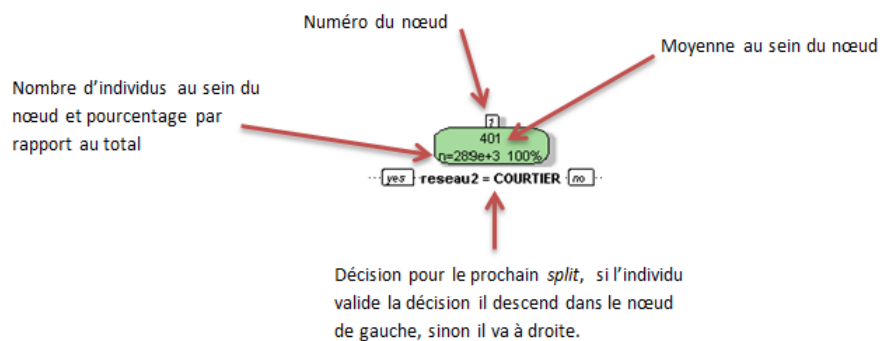
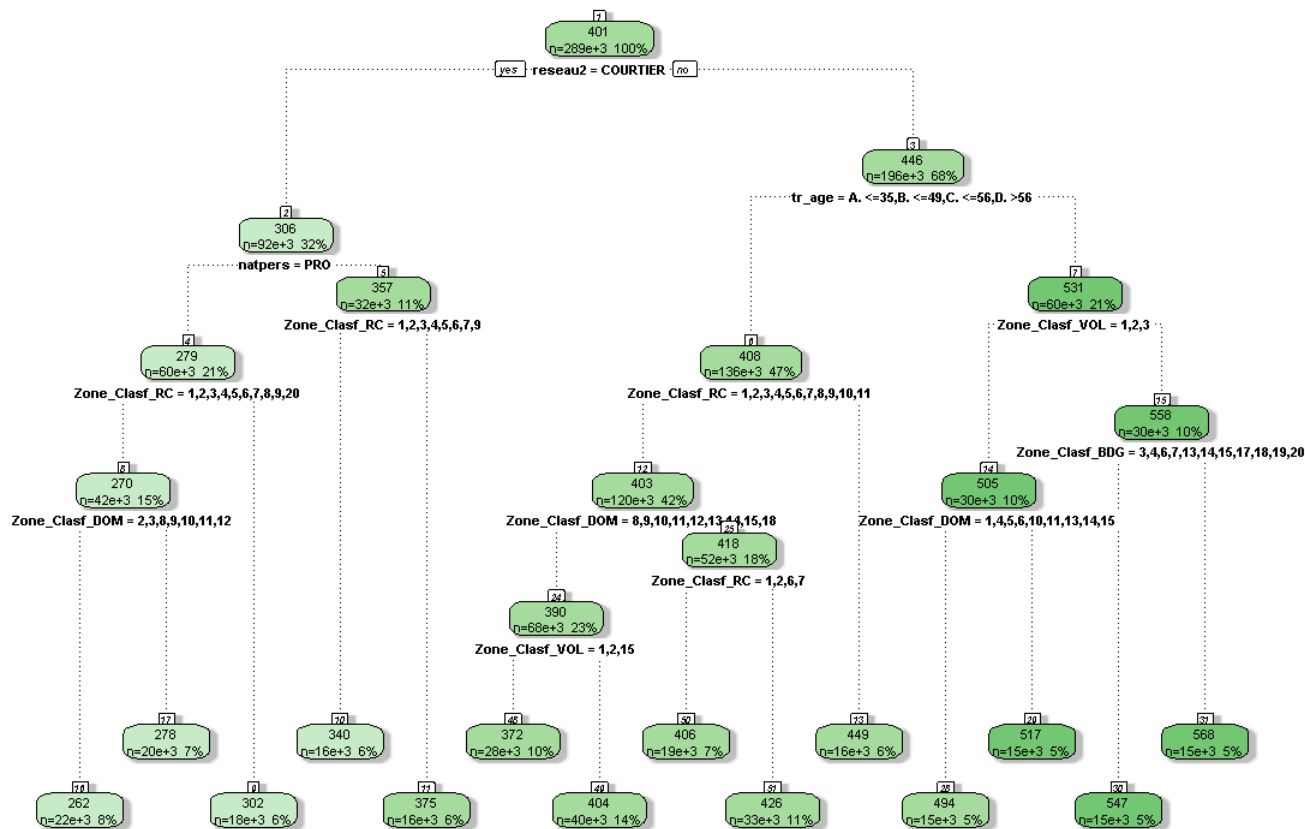


Figure 4.5 – Lecture d'un arbre de décision

Pour le nœud ci-dessus, les clients appartenant au réseau agent vont dans la feuille située à droite. Et les clients appartenant au réseau de courtage vont à gauche de l'arbre.



Rattle 2019-juil.-25 14:30:41 b011pqo

Figure 4.6 – Arbre de régression pour la valeur des produits *Automobile hors E.S.*

L'arbre a une profondeur de 5. Les premières variables prises en compte par l'arbre sont le *réseau de distribution* (reseau2), la *nature du client* (natpers) et l'*âge* (tr_age). Les variables de zoniers interviennent seulement dans les derniers nœuds. Les prédictions vont de 262 à 568, ainsi l'arbre permet de balayer un large intervalle de valeur.

4.3.1.2 ARBRE DE RÉGRESSION POUR LES PRODUITS *Engins Spéciaux*

L'**histogramme 4.7**, ci-dessous, représente la répartition de la valeur des contrats *Engins Spéciaux* hors offres commerciales. La valeur est très concentrée entre 0 et 300 euros.

L'arbre de régression (**figure 4.8**) a une profondeur de 3. Les individus n'ayant pas la modalité *Entreprise* ou *PRO* (i.e. : ceux possédant la modalité *Petites entreprises*) ont des valeurs plus élevées. Toutefois les prédictions vont de 195 et 241 ce qui est un faible intervalle. De plus, les écarts de valeur pour certaines feuilles sont seulement de 10. Ainsi, l'intérêt de cet arbre de régression est assez faible.

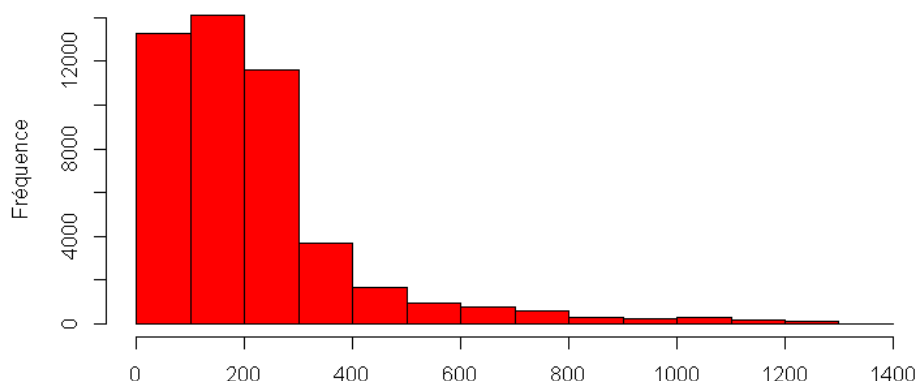


Figure 4.7 – Répartition de la valeur hors offre commerciale des contrats *Engins Spéciaux*

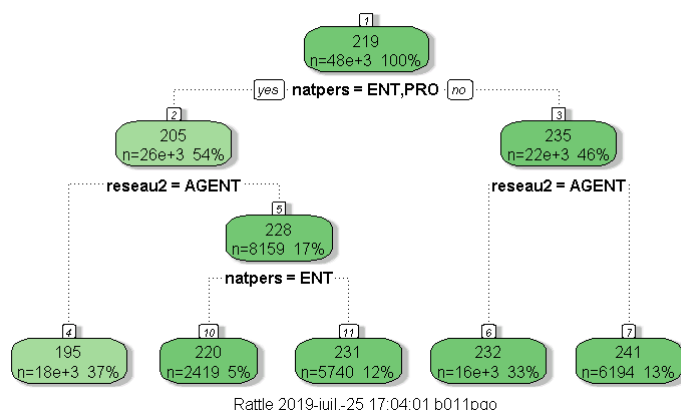


Figure 4.8 – Arbre de régression pour la valeur des produits *Engins Spéciaux*

4.3.1.3 ARBRE DE RÉGRESSION POUR LE PRODUIT *MRH & PNO*

L'**histogramme 4.9**, ci-dessous, représente la répartition de la valeur des contrats *MRH & PNO* hors offres commerciales. La valeur a une distribution très asymétrique.

L'arbre a une profondeur de 5. Les premières variables prises en compte par l'arbre sont l'*âge* (*tr_age*), le *réseau de distribution* (*reseau2*) et la *nature du client* (*natpers*). Les variables de zoniers interviennent seulement dans les derniers nœuds. Les prédictions vont de 158 à 358, ainsi l'arbre permet de balayer un large intervalle de valeur.

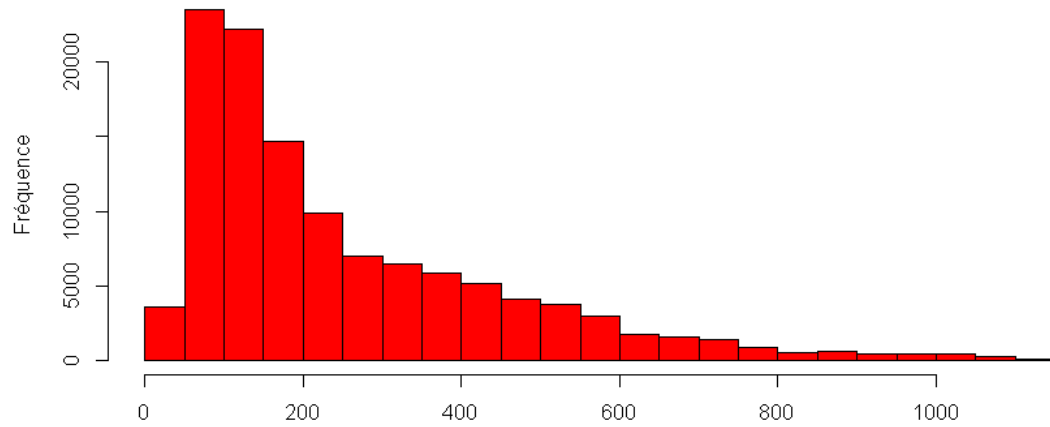
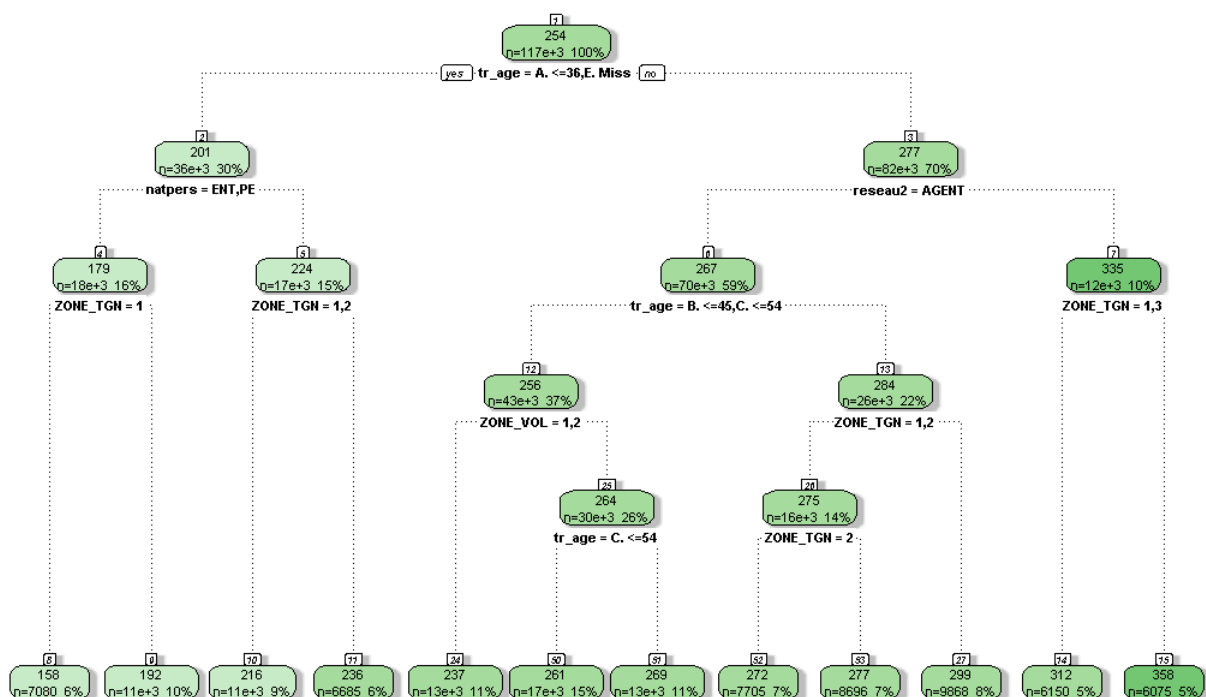


Figure 4.9 – Répartition de la valeur hors offre commerciale des contrats
MRH & PNO



Rattle 2019-juil.-25 15:46:13 b011pqo

Figure 4.10 – Arbre de régression pour la valeur des produits *MRH* & *PNO*

4.3.2 Modélisation de la distribution de la valeur par classe

Une fois l'arbre de régression construit, la distribution de valeur au sein de chaque feuille est observée. Ainsi, nous constatons que la valeur des contrats demeurent très dispersées au sein des feuilles. Cela s'explique par le fait que nous tentons de modéliser la valeur des futurs contrats souscrits. Dès lors, nous ne possédons pas ou très peu de variables expliquant le risque du client et donc le tarif du contrat. C'est pourquoi la valeur reste volatile.

Il est donc décidé de modéliser la distribution des valeurs au sein des feuilles.

Trois lois sont ajustées pour chaque feuille par maximum de vraisemblance :

- la loi Gamma ;
- la loi de Weibull ;
- la loi LogNormale.

La loi Gamma et la loi LogNormale sont choisies car elles sont asymétriques et positives. Enfin, la loi Weibull a été choisie car nous observons certaines queues de distribution lourdes parmi les densités au sein des feuilles (**figure 4.11**).

4.3.2.1 STATISTIQUES D'AJUSTEMENT DE LOIS DE PROBABILITÉ

Trois statistiques mesurant la qualité de l'ajustement entre un échantillon et une loi continue sont présentées dans cette partie. Ces statistiques seront par la suite utilisées pour le choix des lois.

Soit la fonction de répartition empirique de l'échantillon X de taille n :

$$\forall x \in \mathbb{R} \quad F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_i \leq x}$$

Et soit F la fonction de répartition de loi cible.

1. **La statistique de Kolmogorov-Smirnov** s'écrit :

$$D_{KS}(F_n, F) = \max_x (F_n(x) - F(x))$$

Soit l'hypothèse $\mathbb{H}_0$: la statistique $D_{KS}(F_n, F)$ ne dépend pas de F . De plus, pour les échantillons de taille suffisante, il existe le résultat asymptotique :

$$\mathbb{P}_{\mathbb{H}_0}[\sqrt{n}D_{KS}(F_n, F) \leq t] \xrightarrow{n \rightarrow +\infty} 1 - 2 \sum_{k=1}^{\infty} (-1)^{k+1} \exp^{-2k^2 t^2}$$

2. **La statistique de Cramer-von Mises** s'écrit :

$$w^2 = \int_{-\infty}^{+\infty} [F_n(x) - F(x)]^2 dF(x)$$

Cette statistique mesure la moyenne des écarts quadratiques entre les distributions. Elle permet de mesurer l'ajustement sur l'ensemble de la fonction de répartition.

3. **La statistique d'Anderson-Darling** s'écrit :

$$A^2 = n \int_{-\infty}^{+\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} dF(x)$$

Cette statistique mesure la moyenne des écarts quadratiques entre les distributions tout comme la statistique précédente. Toutefois, les écarts sont ici pondérés, les observations présentes dans les queues de distribution ont un poids plus important. Ainsi, cet indicateur permet de mesurer l'ajustement sur les queues de distribution.

Il est donc intéressant d'observer ces trois statistiques simultanément, car elles se complètent.

4.3.2.2 AJUSTEMENT DES LOIS DE DISTRIBUTION DE LA VALEUR DES CONTRATS POUR CHAQUE FEUILLE

Le **graphique 4.11** ci-dessous représente la distribution de la valeur des contrats au sein de chaque feuille. La valeur moyenne prédite par l'arbre de régression ne prend pas en compte la dispersion au sein des feuilles. Il est alors nécessaire d'ajuster des lois de probabilité afin de pouvoir intégrer cette volatilité dans les applications qui suivront.

La distribution de la valeur pour les deux autres arbres de régression est disponible en annexe B.

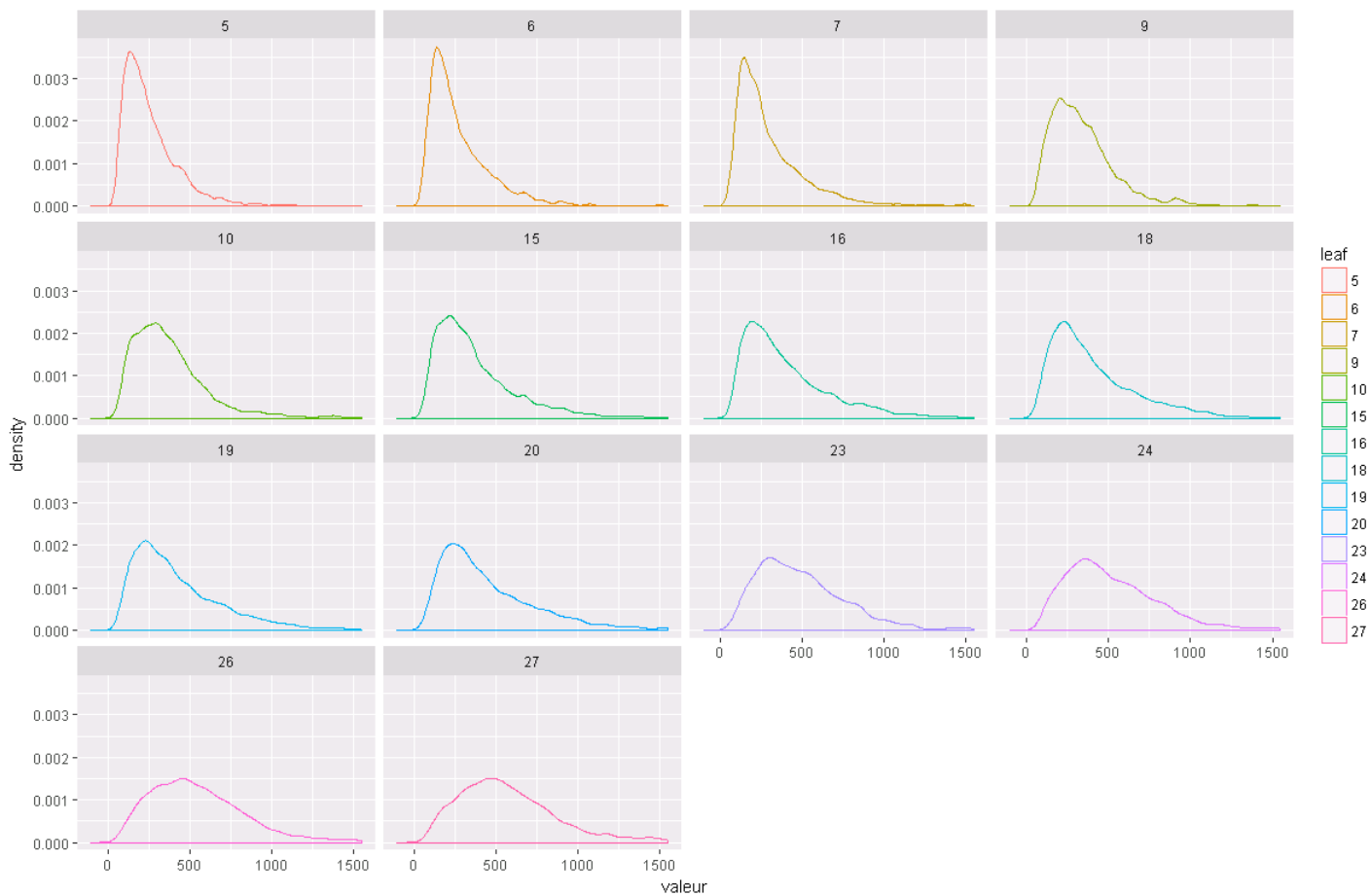


Figure 4.11 – Répartition de la valeur au sein de chaque feuille de l'arbre de régression pour les produits *Automobile hors E.S.*

La démarche de sélection d'une loi de probabilité n'est réalisée que pour une seule feuille dans la suite de cette partie. Les statistiques et graphiques des autres feuilles sont présents en annexe B.

La détermination de la loi de probabilité passe par l'analyse de trois graphiques (la densité, la fonction de répartition et le diagramme quantile-quantile) et des trois statistiques qui ont été présentées dans la *partie 4.2.2.1*.

Intéressons nous à la 9-ième feuille de l'arbre de régression pour les produits *Automobile hors E.S.* Elle correspond au numéro 19 dans les **graphiques 4.11**.

Sur le **graphique 4.12**, ci-dessous, sont représentés les densités, diagrammes quantiles-quantiles et fonctions de répartition des lois ajustées et de la loi empirique. La médiane et la moyenne de la distribution empirique sont aussi représentées car cela permet de constater l'asymétrie importante de la distribution.

Enfin, sur les graphiques suivants, la courbe rouge représente la loi LogNormale, la courbe bleu la loi Gamma et la courbe verte la loi Weibull.

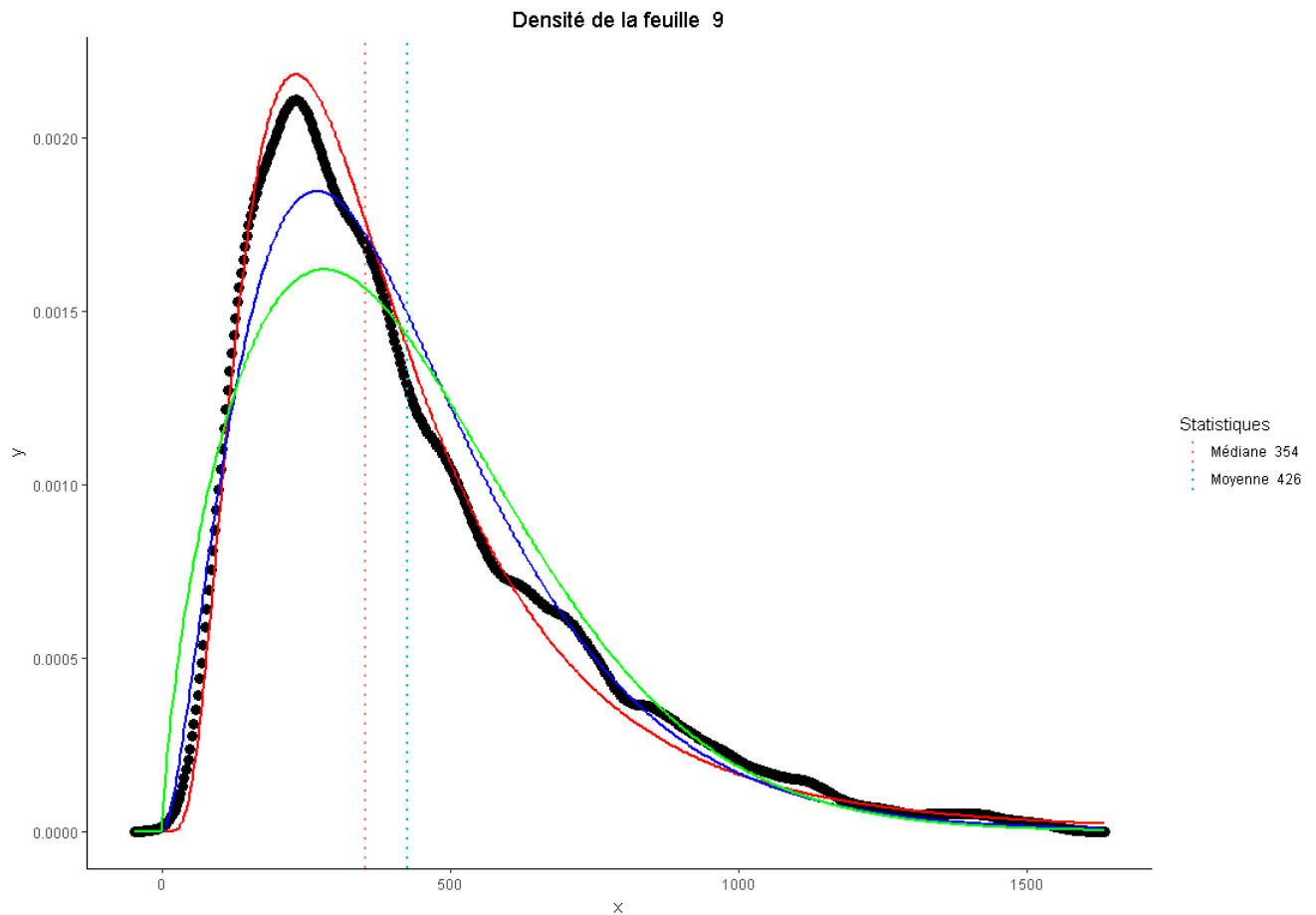


Figure 4.12 – Densités de la empirique et des lois ajustées

Les trois distributions ont une valeur modale proche de la valeur empirique. Cependant, la densité du mode pour les lois de Weibull et Gamma sont bien inférieures à la densité empirique. Cela a un impact sur le reste de la distribution puisque la densité qui n'est pas attribuée au mode va se répartir sur le reste des valeurs.

L'observation du diagramme quantile-quantile (**graphique 4.13**) montre que jusqu'au quantile de valeur 1000, chacune des lois est correctement ajustée. Au-delà de cette valeur, les lois de Weibull et Gamma sont au-dessus de la première bisectrice : cela signifie qu'elles sous-estiment la queue de distribution. Concernant la loi LogNormale, elle sur-estime la queue de distribution.

Ces conclusions sont à modérer car le graphique des fonctions de répartition (**graphique 4.14**) montre que les observations de la valeur 1000 correspondent à moins de 5 % de la distribution.

De plus, le **graphique 4.14** nous apprend aussi que les répartitions des lois Weibull et Gamma sont moins bien ajustées. En effet, avant la valeur 200, leur répartition est supérieure à la répartition empirique et entre la valeur 200 et 500, leur répartition est inférieure à la répartition empirique. Cela est dû au fait que la densité modale, pour ces lois, est très inférieure à la densité empirique, ce qui a un impact sur le reste de la distribution.

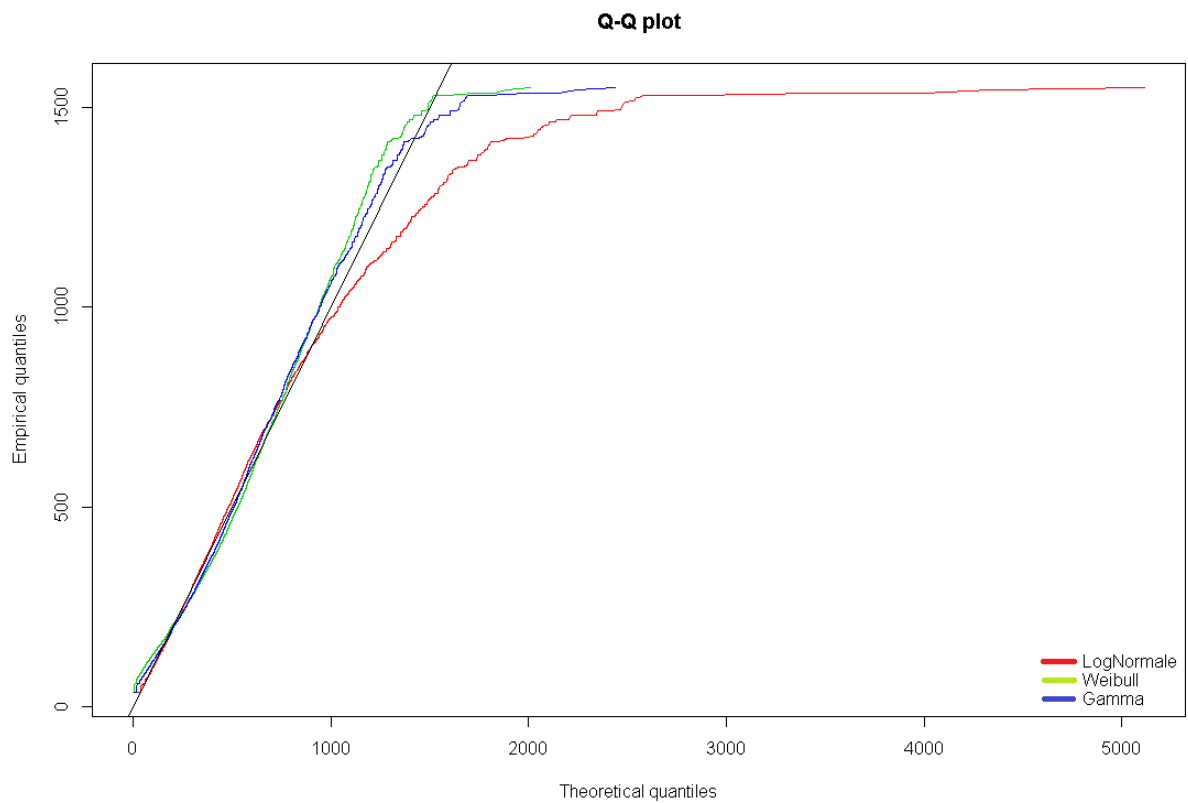


Figure 4.13 – Diagramme quantile-quantile

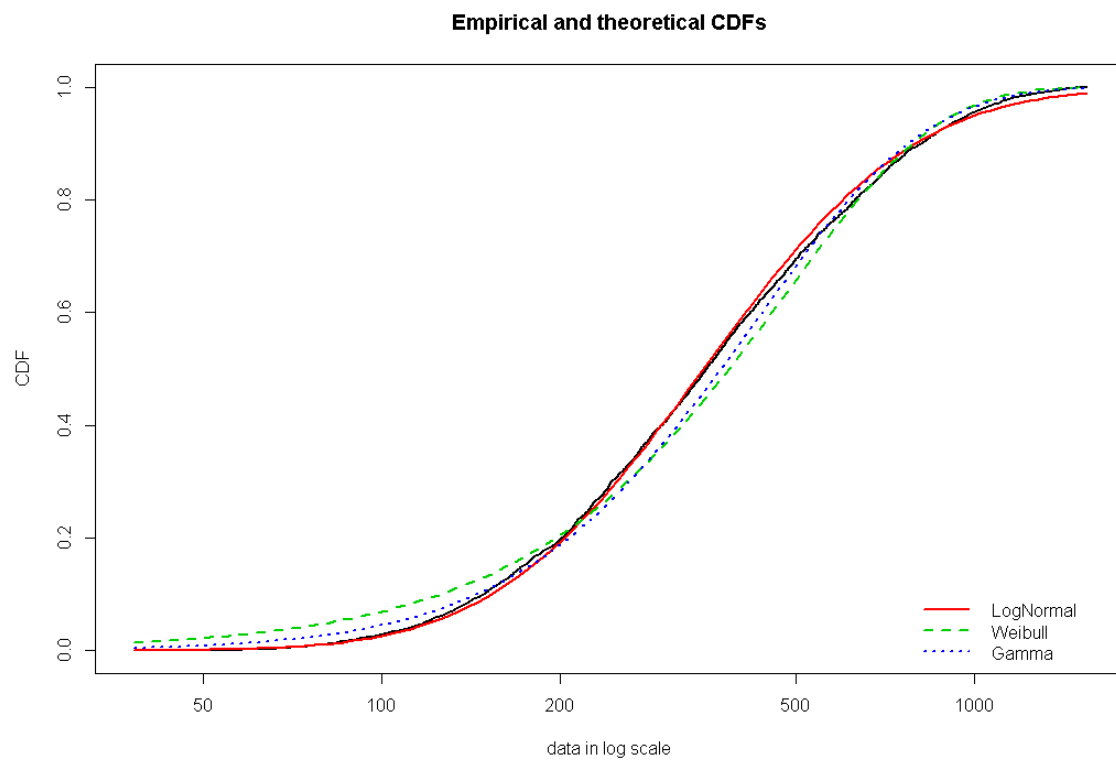


Figure 4.14 – Fonctions de répartition empirique et des lois ajustées

Goodness-of-fit statistics			
	LogNormal	weibull	Gamma
Kolmogorov-Smirnov statistic	0.02466384	0.05511601	0.03888582
Cramer-von Mises statistic	4.09193005	33.32168123	16.24234316
Anderson-Darling statistic	31.42270740	219.73433894	97.06709769

Figure 4.15 – **Statistiques d’ajustement**

Enfin, les statistiques d’ajustement (**tableau 4.15**) confirment que la loi LogNormale est, des trois lois, celle qui ajuste le mieux la distribution empirique des données.

Le **tableau 4.16** résume les différentes lois qui ont été ajustées à chaque feuille de chaque arbre.

Produit	Feuille	Loi de distribution	Paramètres de la loi			
			Shape	Rate	Moyenne	Ecart-type
Auto hors E.S.	1	Log-Normale			5,38	0,61
	2	Log-Normale			5,43	0,63
	3	Log-Normale			5,50	0,64
	4	Gamma	3,28	0,01		
	5	Log-Normale			5,76	0,59
	6	Log-Normale			5,72	0,64
	7	Log-Normale			5,80	0,65
	8	Log-Normale			5,81	0,64
	9	Log-Normale			5,86	0,64
	10	Log-Normale			5,91	0,64
	11	Gamma	3,39	0,01		
	12	Gamma	3,48	0,01		
	13	Gamma	3,58	0,01		
	14	Gamma	3,60	0,01		
Engins Spéciaux	1	Gamma	1,77	0,01		
	2	Log-Normale			5,12	0,72
	3	Log-Normale			5,11	0,82
	4	Gamma	1,62	0,01		
	5	Gamma	1,58	0,01		
MRH & PNO	1	Log-Normale			4,82	0,64
	2	Log-Normale			5,01	0,66
	3	Log-Normale			5,12	0,71
	4	Log-Normale			5,19	0,75
	5	Log-Normale			5,19	0,76
	6	Log-Normale			5,25	0,8
	7	Log-Normale			5,32	0,77
	8	Log-Normale			5,32	0,77
	9	Log-Normale			5,34	0,78
	10	Log-Normale			5,43	0,76
	11	Log-Normale			5,47	0,77
	12	Gamma	2,18	0,01		

Figure 4.16 – Tableau récapitulatif des lois de probabilité choisies pour chaque feuille

4.3.3 Illustration du modèle de rentabilité potentielle

Afin d'illustrer le modèle de rentabilité potentielle qui vient d'être construit, celui-ci est appliqué sur deux individus ayant des appétences pour des produits différents (cf. **tableau 4.3**).

Le **graphique 4.17** représente la distribution de valeur en cas de souscription d'un individu appartenant à la classe "*PRO - IARD 10*" (en *vert*) et "*Non PRO - Mixte*" (en *rouge*). Détaillons de plus les feuilles dans lesquelles les deux individus ont été prédits :

- Concernant l'individu *rouge* :
 - la valeur moyenne pour le produit *Automobile hors E.S.* prédite est de 568, soit la 14-ième feuille ;
 - la valeur moyenne pour le produit *Engins Spéciaux* prédite est de 232, soit la 4-ième feuille ;

- la valeur moyenne pour le produit *MRH & PNO* prédite est de 192, soit la 2-ième feuille.
- Concernant l'individu *vert* :
 - la valeur moyenne pour le produit *Automobile hors E.S.* prédite est de 262, soit la 1-ère feuille ;
 - la valeur moyenne pour le produit *Engins Spéciaux* prédite est de 231, soit la 3-ième feuille ;
 - la valeur moyenne pour le produit *MRH & PNO* prédite est de 216, soit la 3-ième feuille.

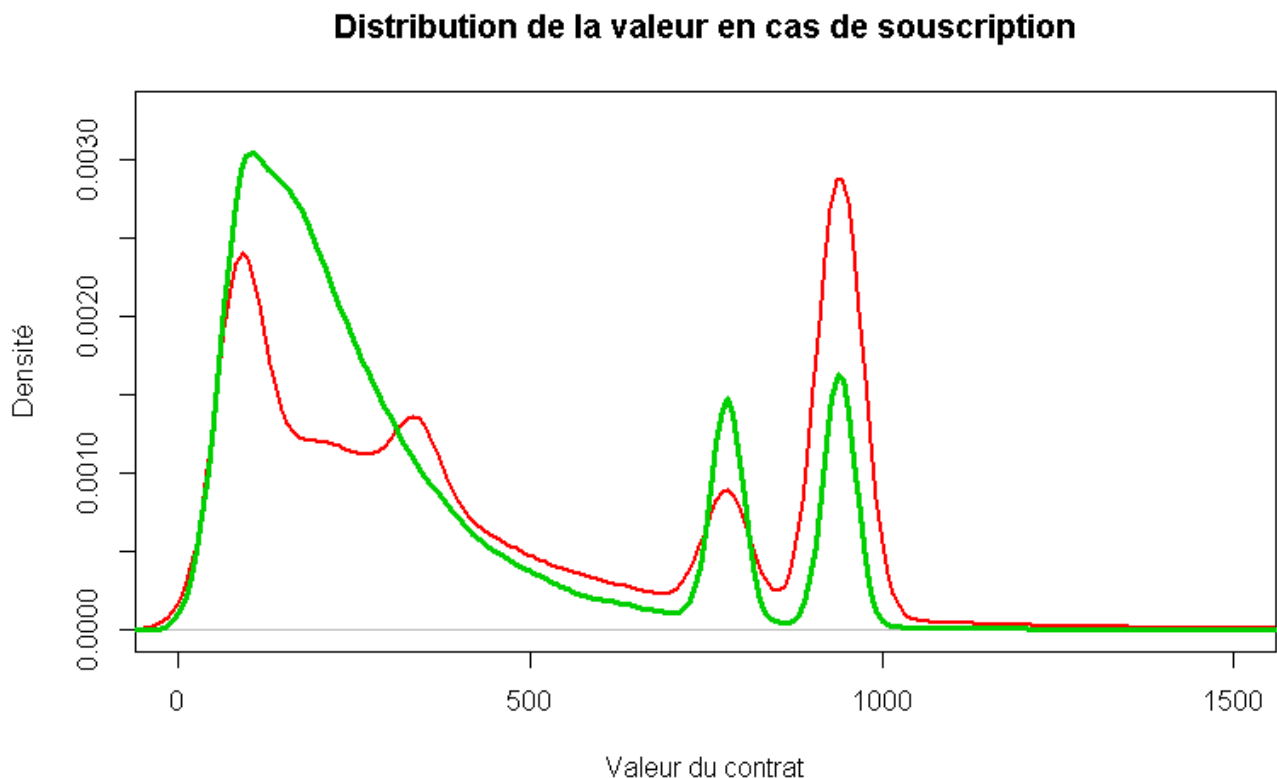


Figure 4.17 – **Densité de la valeur en cas de souscription**

L'individu *rouge* présente un pic de densité autour de la valeur 900. Cela s'explique par le fait qu'il possède 23 % de chance de souscrire un contrat '*Autre IARD*'. De même l'individu présente un pic autour de la valeur 700 car il possède 10 % de chance de souscrire un contrat de prévoyance (cf. **tableau 4.3**).

Expliquons maintenant la partie gauche de la distribution :

- L'individu *vert* a été prédit dans la classe qui présente la plus faible valeur pour le contrat *Automobile hors E.S.*, de plus, il présente 52 % de chance de souscrire un contrat *Automobile hors E.S.* Cela explique pourquoi la densité est très élevée entre 100 et 350 de valeur.

- L'individu *rouge* présente des probabilités peu élevées de souscrire des produits *MRH* & *PNO* (6 %) et *Engins Spéciaux* (9 %). De plus, concernant le produit *Automobile hors E.S.*, il a été prédit dans la classe présentant la plus forte valeur pour ce contrat (568 euros de valeur en moyenne).

Ces deux points expliquent d'une part pourquoi la densité de la valeur est plus faible à gauche de la distribution (entre 200 et 400) et d'autre part pourquoi la densité de la valeur est supérieure à celle de l'individu vert au milieu de la distribution (entre 400 et 700).

- La distribution de l'individu *rouge* présente un pic autour de 100 de valeur, cela s'explique par le fait que sa probabilité de souscrire un contrat *MRC* est de 13 %.

Enfin, le **tableau 4.18** présente les quartiles de la distribution de valeur en cas de souscription de chacun des individus :

	1 ^{er} quartile	Médiane	Moyenne	3 ^{ème} quartile
Individu vert	134	231	347	460
Individu rouge	166	362	488	940

Figure 4.18 – **Quartiles et moyennes des distributions**

En résumé, le potentiel de valeur en cas de souscription de l'individu *vert* est inférieur de près de 140 euros en moyenne par rapport à l'individu *rouge*.

4.3.4 Conclusion

L'estimation de la distribution de la valeur en cas de souscription d'un nouveau contrat permet de réduire l'incertitude puisque cette incertitude est maintenant quantifiée au sein d'une loi de probabilité.

De plus, la modélisation de la fonction de répartition permet de prendre en compte l'asymétrie de la distribution de la valeur en cas de souscription.

Enfin, l'utilisation de fonction de répartition permet d'aborder la valeur potentielle de façon complètement probabiliste. En effet, seulement la souscription d'un nouveau contrat est probabilisée dans le modèle actuellement en place.

La probabilité de multi-équipement a été modélisée dans le **chapitre 3** et la distribution de valeur en cas de souscription vient d'être modélisée dans ce chapitre. À partir de ces deux indicateurs, la valeur potentielle du client est construite. La moyenne de la distribution de valeur est multipliée par la probabilité d'équipement.

Chapitre 5

La valeur potentielle et son application

Dans ce chapitre, nous proposons une application de la valeur potentielle. Cette application concerne la distribution d'offres commerciales (ou rabais) à des prospects qui sont considérés comme présentant du potentiel.

Notre but est de voir s'il est possible d'optimiser la valeur d'un portefeuille de prospects.

La **partie 5.1** détaille l'application qui va être développée. Ensuite la base de donnée qui va être utilisée sera présentée (**partie 5.2**). Enfin, les résultats de l'application seront analysés (**partie 5.3**).

5.1 Application à la politique de rabais

5.1.1 Idée de l'application

La valeur potentielle permet :

- d'identifier des clients possédant une forte probabilité de se multi-équiper ;
- de connaître la distribution de valeurs en cas de souscription.

L'idée centrale de cette application est d'identifier les prospects à potentiel et de leur accorder un rabais sur la prime de leur premier contrat afin qu'ils entrent en portefeuille. Ce rabais est indexé sur la valeur des futurs contrats du client.

Ainsi, nous sommes prêt à dégrader la valeur du premier contrat des clients présentant du potentiel, pour qu'ils entrent en portefeuille et qu'ils réalisent leur potentiel. Toutefois une borne maximale de rabais sur la prime est fixée à 50 % de la prime du premier contrat. Ce choix est fait pour ne pas trop dégrader le chiffre d'affaire du produit d'appel.

Différents scénarios seront réalisés dans lesquels d'une part le pourcentage de clients bénéficiant de rabais sera amené à varier. Par exemple, le premier scénario n'accorde un rabais qu'aux 20 % de clients présentant la probabilité de multi-équipement la plus forte, alors que le scénario suivant concerne les 30 % de clients avec la plus forte probabilité.

D'autre part, le montant du rabais accordé en fonction de la valeur potentielle sera aussi amené à varier. Par exemple, un scénario dégradera la valeur du premier contrat de 10 % de la valeur potentielle, alors qu'un autre scénario la dégradera de 20 % de la valeur potentielle.

Remarque : Une perte de 1 euro de valeur n'est pas équivalent à un rabais sur la prime de 1 euro ! La valeur du contrat étant projetée, le rabais de 1 euro a pour effet de diminuer tous les flux futurs, ainsi la perte en valeur est plus importante que le rabais accordé. Nous expliquons dans l'annexe C quel est le lien entre le rabais sur la prime et la valeur du contrat.

La métrique utilisée pour comparer les scénarii est la valeur de l'ensemble du portefeuille de prospects étudié au bout d'un an (i.e. : la somme de la valeur de chaque client). Nous nous plaçons au bout d'un an car la probabilité de multi-équipement a été calculée sur un tel horizon.

Au sein de chaque scénario, 400 simulations stochastiques sont réalisées afin d'avoir une estimation précise de la répartition de la valeur du portefeuille au bout d'un an. Si un client est amené à se multi-équiper dans une simulation, la valeur du contrat est tirée selon la loi de distribution associée au client qui a été calculée dans le **chapitre 4**.

5.1.2 Hypothèse de l'application

Le taux de transformation est la probabilité qu'une affaire se réalise. Il correspond à la proportion de clients qui ont réellement souscrit un contrat parmi tous les clients entrés dans une agence.

Le taux de transformation est actuellement de 50 % à Generali. Ce taux est considéré dans la suite comme étant le taux de base pour nos simulations (i.e. : le taux dans le cas où il n'y a pas d'offre commerciale).

Dans le cadre de nos simulations, le taux de transformation sera supérieur à 50 % lorsqu'un client bénéficie d'un rabais. **L'hypothèse selon laquelle le taux de transformation est le même pour tous les clients bénéficiant d'un rabais est supposée.** Il s'agit d'une hypothèse très forte et une étude sur l'élasticité au rabais des prospects devrait être réalisée afin de l'affiner.

De plus, il n'est pas déraisonnable d'émettre l'hypothèse que le taux de transformation évolue de façon croissante en fonction du rabais accordé. C'est pourquoi un taux de transformation raisonnablement faible est choisi lorsque la perte de valeur est faible et inversement.

5.2 La base de données utilisée

La base de données est composée de **4500 affaires nouvelles** réalisées en *Automobile* durant l'année 2018. Ces affaires nouvelles ont toutes été réalisées par des nouveaux clients. Ainsi notre étude traitera de l'attribution d'offres commerciales à ces 4500 clients comme s'il s'agissait de prospects se présentant en agence pour souscrire un contrat *Automobile*.

Les variables composant la base de données sont :

- la prime hors rabais associée au contrat ;
- la valeur associée à une telle prime.

Le **graphique 5.1** représente la répartition de la prime et de la valeur de la base des prospects. La ligne bleu indique la moyenne de l'échantillon.

La moyenne des prime est de 890 euros et la moyenne des valeur est de 350 euros.

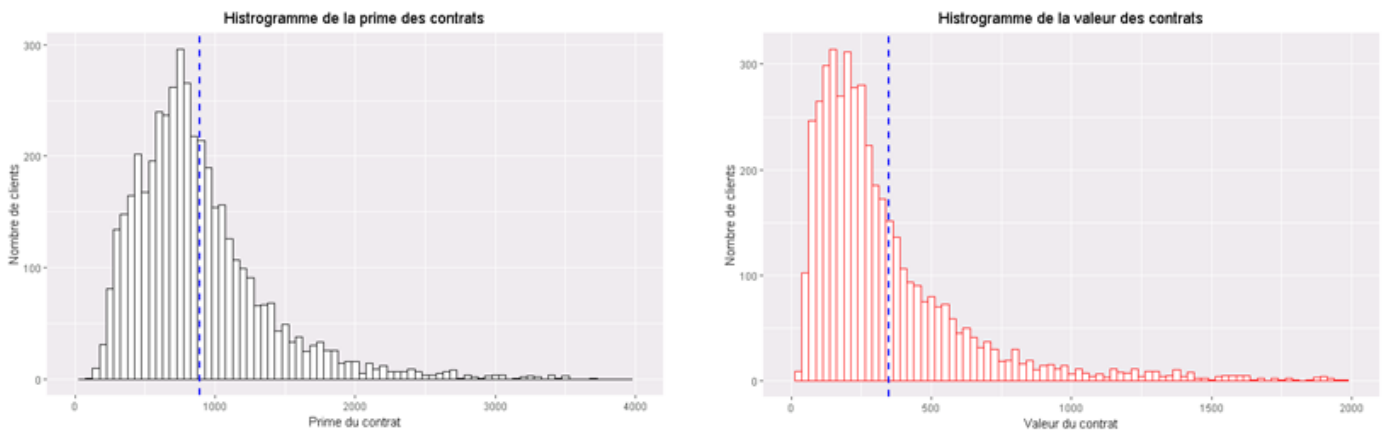


Figure 5.1 – **Histogramme de la prime et de la valeur des contrats**

Ensuite la valeur potentielle de chaque individu est calculée (probabilité et distribution de valeur).

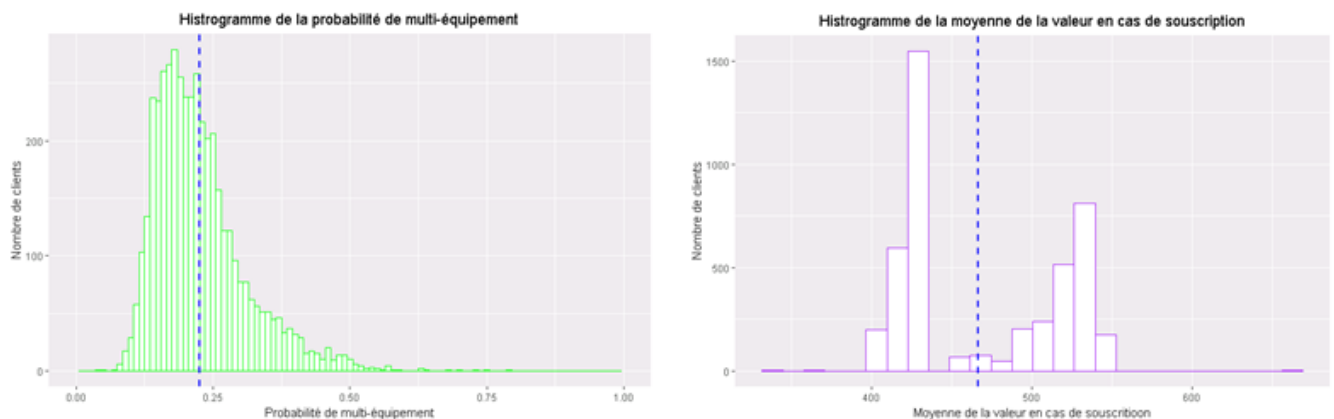


Figure 5.2 – **Histogramme de la probabilité prédite et de la valeur moyenne en cas de souscription des clients**

Lecture :

- La moyenne des probabilités de multi-équipement est de 23 %, cette valeur est très grande par rapport aux probabilités trouvées dans le **chapitre 3**. Cela s'explique par le fait que notre base n'est composée que de clients étant entrés par un contrat *Automobile*. Or, il a été montré qu'il s'agit d'une population avec un potentiel très fort (cf. **partie 2.2.2**).
- La valeur en cas de souscription du client est très dépendante de la classe à laquelle appartient le client (cf. **figure 4.3**). Cela explique pourquoi la distribution de la moyenne de la valeur en cas de souscription est très concentrée entre 400 et 435 et entre 500 et 550.

5.3 Résultats de l'application

5.3.1 Scénario témoin

Dans ce scénario, la politique de rabais n'est pas appliquée. Il s'agit du *Scénario témoin*. Ainsi, tous les prospects de la base ont un taux de transformation de 50 %.

Le **tableau 5.3** résume les statistiques de la répartition de la valeur du portefeuille au bout d'un an.

	Valeur totale au bout d'un an	Valeur du premier contrat	Valeur liée au multi-équipement
Min.	930	736	190
1 ^{er} quartile	989	769	215
Médiane	1003	781	222
Moyenne	1003	780	222
3 ^{ème} quartile	1018	791	230
Max.	1064	830	257

Figure 5.3 – Statistiques descriptives liées à la valeur du portefeuille (en milliers) à la fin de la première année pour le *Scénario témoin*

Le ratio moyen entre la valeur liée au multi-équipement et la valeur totale est de 22 %. En moyenne, la valeur du portefeuille au bout d'un an est d'un million d'euros, ce qui correspond à 450 euros en moyenne par client en portefeuille :

$$\frac{1003}{50\% \times 4500} = 0.45 \text{ k€ de valeur}$$

Où : 4500 est le nombre de prospects de la base et 50 % est le taux de transformation.

La répartition des gains est assez symétrique puisque les écarts entre les quartiles et la médiane sont respectivement de 14 000 (entre le premier quartile et la médiane) et de 15 000 (entre le troisième quartile et la médiane). De plus, si l'on compare l'écart interquartile de la valeur du premier contrat (23 000) et de la valeur liée au multi-équipement (16 000) nous constatons que les gains liés au multi-équipement sont moins volatiles. Cela s'explique par le fait que la probabilité moyenne de multi-équipement est plus de deux fois plus faible que le taux de transformation (23 % contre 50 %).

5.3.2 Applications des scénarii accordant des rabais

5.3.2.1 Scénario 1

Dans le premier scénario testé (*Scénario 1*), 10 % des prospects se voient offrir une offre commerciale équivalente à 10 % de leur valeur potentielle moyenne. Le taux de transformation des clients bénéficiant d'une telle offre est de 55 %. Une telle offre engendre un rabais moyen de 18 euros seulement, c'est pourquoi nous avons choisi un taux de transformation relativement proche des 50 % du scénario témoin.

	Valeur totale au bout d'un an	Valeur du premier contrat	Valeur liée au multi-équipement
Min.	915	724	191
1 ^{er} quartile	991	766	220
Médiane	1004	777	228
Moyenne	1005	778	228
3 ^{ème} quartile	1019	789	235
Max.	1065	824	266

Figure 5.4 – Statistiques descriptives liées à la valeur du portefeuille (en milliers) à la fin de la première année pour le *Scénario 1*

Le gain de valeur au bout d'un an entre le *Scénario témoin* et le *Scénario 1* est **en moyenne de 2 000 euros (soit 0,2 %)**.

Explication du gain de valeur en moyenne :

- une perte d'environ 3 000 euros sur la valeur des premiers contrats bénéficiant de l'offre. Cela est dû aux rabais accordés sur les primes des clients à potentiel de multi-équipement ;
- un gain de 5 000 euros sur la valeur des contrats issus du multi-équipement. Cela signifie bien que plus de clients sont multi-équipés.

De plus, en comparant les écarts interquartiles, nous constatons que la volatilité de la distribution des gains totaux baisse légèrement (28 000 contre 29 000 euros si l'on regarde l'écart interquartile). Cela est dû à la baisse de la volatilité des gains liés au multi-équipement (15 000 contre 16 000 euros).

Le scénario qui vient d'être testé augmente la valeur du portefeuille, mais **les gains sont toutefois marginaux**. Le second scénario est plus agressif : les offres commerciales sont plus importantes et le nombre de clients en bénéficiant est plus large aussi.

5.3.2.2 *Scénario 2*

Dans le *Scénario 2*, 50 % des prospects bénéficient d'un rabais, la baisse de valeur sur le premier contrat souscrit correspond à 30 % de la valeur moyenne en cas de souscription et le taux de transformation associé est de 75 %.

L'augmentation du pourcentage de prospects bénéficiant d'une offre commerciale s'explique par le fait que notre base de données contient beaucoup de clients présentant un potentiel de multi-équipement très élevé. Ainsi, nous pouvons nous permettre d'étendre autant le périmètre d'attribution des offres commerciales.

La baisse de valeur de 30 % correspond à un rabais moyen sur la prime de 55 euros.

	Valeur totale au bout d'un an	Valeur du premier contrat	Valeur liée au multi-équipement
Min.	996	715	253
1 ^{er} quartile	1038	746	285
Médiane	1050	755	295
Moyenne	1051	755	295
3 ^{ème} quartile	1064	763	305
Max.	1108	790	336

Figure 5.5 – Statistiques descriptives liées à la valeur du portefeuille (en milliers) à la fin de la première année pour le *Scénario 2*

Le gain de valeur au bout d'un an entre le *Scénario témoin* et le *Scénario 2* est **en moyenne de 48 000 euros (soit 4,8 %)**.

Explication du gain de valeur en moyenne :

- une perte de 23 000 euros sur la valeur des contrats bénéficiant de l'offre ;
Les rabais sont plus importants et attribués plus largement que dans le *Scénario 1*.
- un gain de 73 000 euros sur la valeur des contrats issus du multi-équipement.

L'observation des écarts interquartiles nous montre que la volatilité de la valeur totale a baissé (l'écart interquartile vaut 26 000 euros). Cela s'explique par le fait que la volatilité de la valeur du premier contrat a chuté : en effet l'écart interquartile est passé de 23 000 à 17 000 euros. Or, la valeur du premier contrat représente une grande partie de la valeur totale, d'où la réduction de la dispersion de celle-ci.

La volatilité de la valeur du premier contrat baisse car le taux de transformation pour les clients bénéficiant d'une offre est de 75 %. L'incertitude concernant la transformation de ces clients est donc diminuée, ce qui mécaniquement fait baisser l'incertitude totale sur la valeur du premier contrat.

Le ratio moyen entre la valeur liée au multi-équipement et la valeur totale est de 28 %. De plus, la valeur moyenne par client est désormais de 375 euros :

$$\frac{1054}{(50\% \times 75\% + 50\% \times 50\%) \times 4500} = 0.375 \text{ k€ de valeur}$$

Grâce aux rabais sur les primes, plus de clients sont entrés en portefeuille, d'où la baisse de la valeur moyenne.

Intéressons nous à la valeur moyenne des clients :

- La valeur moyenne des clients qui ont reçu une offre commerciale est de 384 euros.

Les clients qui bénéficient d'une offre commerciale ont une probabilité moyenne de se multi-équiper de 28 %. Cela signifie que plus de 70 % de ces clients ne se multi-équipent pas et gardent donc une valeur dégradée suite au rabais sur la prime. C'est pourquoi la valeur est en moyenne inférieure à 450 euros (450 euros étant la moyenne de valeur dans le *Scénario témoin*).

- La valeur moyenne des clients qui n'ont pas reçu d'offre commerciale est de 358 euros.

Les clients qui n'ont pas reçu d'offres sont les clients qui présentent un potentiel faible de multi-équipement. C'est pourquoi ils présentent une valeur moyenne plus faible.

5.3.2.3 Scénario 3

Le *Scénario 2* étant générateur de gains et présentant un gain sur la valeur du portefeuille de 4,8 % en moyenne, il est décidé de réitérer ce scénario en accordant cette fois des offres commerciales à l'intégralité du portefeuille. Le taux de transformation est le même que précédemment (75 %).

	Valeur totale au bout d'un an	Valeur du premier contrat	Valeur liée au multi-équipement
Min.	962	659	292
1 ^{er} quartile	1016	689	322
Médiane	1029	697	332
Moyenne	1029	697	332
3 ^{ème} quartile	1043	706	342
Max.	1089	726	385

Figure 5.6 – Statistiques descriptives liées à la valeur du portefeuille (en milliers) à la fin de la première année pour le *Scénario 3*

Ce scénario augmente la valeur totale par rapport au *Scénario témoin*. Toutefois, le gain en valeur est inférieur au *Scénario 2* : il n'est que de **26 000 euros (soit 2.6 %)**.

Explication du gain de valeur en moyenne :

- une perte de 83 000 euros sur la valeur des contrats bénéficiant de l'offre ;
- un gain de 110 000 euros sur la valeur des contrats issus du multi-équipement.

Dans ce scénario, la part de la valeur liée au multi-équipement par rapport à la valeur totale est de 32 %.

Si l'on effectue le rapport entre les gains et les pertes :

$$\frac{110000}{83000} = 1,3$$

À titre de comparaison, ce ratio vaut 2,9 pour le **Scénario 2**. Ce ratio permet de déterminer la quantité de valeur créée suite au multi-équipement pour 1 euro de valeur perdu sur le premier contrat. Ainsi, même si le **Scénario 3** gère des gains, la politique est peu efficace selon cet indicateur.

5.3.2.4 Scénario 4

Dans ce scénario, nous décidons d'accorder un rabais sur la prime à 80 % des clients. De plus, deux rabais sont proposés :

- les 40 % des clients ayant la probabilité la plus élevée se voient attribuer un rabais correspondant à une baisse de 30 % de leur valeur moyenne en cas de souscription (cela correspond à un rabais moyen de 55 euros) ;
- les 40 % des clients suivants se voient attribuer un rabais correspondant à une baisse de 10 % de leur valeur moyenne en cas de souscription (cela correspond à un rabais moyen de 19 euros).

Nous souhaitons par cette politique favoriser les clients à fort et moyen potentiel, en leur proposant un rabais en fonction de leur potentiel.

	Valeur totale au bout d'un an	Valeur du premier contrat	Valeur liée au multi-équipement
Min.	974	699	241
1 ^{er} quartile	1026	738	283
Médiane	1040	747	292
Moyenne	1039	747	292
3 ^{ème} quartile	1052	756	300
Max.	1095	788	330

Figure 5.7 – **Statistiques descriptives liées à la valeur du portefeuille (en milliers) à la fin de la première année pour le *Scénario 4***

La valeur totale du portefeuille augmente **en moyenne de 36 000 euros (soit 3,6 %)**.

Explication du gain de valeur en moyenne :

- une perte de 33 000 euros sur la valeur des contrats bénéficiant de l'offre.
- un gain de 69 000 euros sur la valeur des contrats issus du multi-équipement.

Dans ce scénario, la part de la valeur liée au multi-équipement par rapport à la valeur totale est de 28 %.

Intéressons nous à la valeur moyenne des clients :

- La valeur moyenne des clients qui ont reçu la plus forte réduction commerciale est de 400 euros.
- La valeur moyenne des clients qui ont reçu la seconde réduction est de 344 euros.
- La valeur moyenne des clients qui n'ont pas reçu de réduction est de 335.

Le ratio entre les gains et les pertes est de 2,1, c'est-à-dire que pour un euro de valeur dégradée, la valeur liée au multi-équipement augmente de 2,1 euros.

Enfin, l'écart interquartile est de 26 000 euros pour la valeur totale. Il s'agit du même écart que pour le **Scénario 2**.

5.4 Conclusion

L'application réalisée dans ce chapitre utilise le modèle qui a été développé dans ce mémoire. Le modèle a été construit à partir des clients en portefeuille. Toutefois, nous avons vu que le taux de multi-équipement dans les bases de données était peu élevé (6,1 %) par rapport au potentiel réel des clients professionnels, car beaucoup de clients en portefeuille se sont déjà multi-équipés. Ainsi, nous constatons que le modèle est amené à sous-estimer le taux de multi-équipement des prospects. En effet, le taux moyen prédit dans notre simulation est de 23 % alors que le taux moyen réel de multi-équipement des clients entrant par l'*Automobile* est de 28 % (cf. **tableau 2.4**). Dans ce raisonnement, les gains de valeurs observés dans l'application sont alors aussi sous-estimés.

En définitive les simulations qui ont été réalisées montrent qu'il est possible d'optimiser l'attribution des rabais sur les primes des clients en prenant en compte leur valeur potentielle. À l'heure actuelle, Generali déploie une politique de rabais visant d'une part à élargir le portefeuille et d'autre part à optimiser la valeur du client. Le travail qui vient d'être mené sur la valeur potentielle et les offres commerciales pourrait être utilement appliqué au reste de la politique de souscription de la compagnie.

Toutefois, il est important d'avoir à l'esprit les limites des simulations réalisées. Tout d'abord, comme précisé au début de cette partie, nous appliquons le modèle qui a été entraîné sur le portefeuille à des prospects. De plus, les taux de transformation qui ont été appliqués constituent une hypothèse forte qu'il conviendrait d'analyser dans une étude complémentaire. Dans le cadre de ce travail, leur détermination a fait l'objet d'un échange et d'une concertation avec l'équipe.

Conclusion

Au cours de ce mémoire, il a été montré que certains groupes de clients professionnels présentent un fort potentiel au multi-équipement. Nous avons notamment identifié que les clients entrant dans le portefeuille par les produits *Automobile* ou *MRH & PNO* ont plus d'une chance sur quatre de s'équiper d'un autre contrat durant l'année.

De plus, le modèle de valeur potentielle qui a été développé permet de prendre en compte l'appétence au multi-équipement d'un client sur tous les produits, contrairement au modèle actuel de Generali qui modélise seulement quelques scores d'appétence. D'autre part, la répartition de la valeur du nouveau contrat a été modélisée : cela permet de prendre en compte l'asymétrie de valeur entre les contrats. La modélisation de la distribution de la valeur permet aussi de capter et de quantifier l'incertitude inhérente à une telle modélisation dans la mesure où nous ne connaissons pas encore le risque associé au client.

L'application du modèle de valeur potentielle dans l'optimisation des offres commerciales a montré qu'il est possible d'augmenter la valeur totale du portefeuille en ciblant les clients à potentiel.

Enfin, nous devons signaler l'importance de l'hypothèse des taux de transformation qui sont utilisés. Ils conditionnent, en effet, en grande partie les résultats de l'application. Nous avons pris des taux raisonnables dans la mesure où ils évoluent de façon croissante en fonction du rabais accordé. Cependant une étude entre l'élasticité du taux de transformation et du rabais permettrait d'affiner les résultats de l'application.

Bibliographie

- **Azencott Ch.-A.**, « Évaluez un algorithme de classification qui retourne des scores » [en ligne], disponible sur <https://openclassrooms.com/fr/courses/4297211-evaluez-et-ameliorez-les-performances-dun-modele-de-machine-learning/4308261-evaluez-un-algorithme-de-classification-qui-retourne-des-scores> [consulté le 8 septembre 2019].
- **Boyer Claire**, « Apprentissage statistique », ISUP, 2018.
- **Charpentier Arthur**, *Mathématiques de l'assurance non-vie*, tomes 1 et 2, Economica, 2004.
- **Fontaine Léonard**, « La modélisation de la valeur contrat (PNPV) par la refonte du modèle de résiliation », Mémoire d'actuariat, ISFA, 2011.
- **Friedman J.H.**, *Greedy function approximation : a gradient boosting machine*, Annals of Statistics, 29, p. 1189–1232, 2001.
- **Gannaz Irène**, « Introduction à la statistique » [en ligne], INSA de Lyon, 2016, disponible sur http://math.univ-lyon1.fr/homes-www/gannaz/Cours/cours_stat.pdf [consulté le 8 septembre 2019].
- **Garotin Mathilde**, « Market-consistent embedded value », ISUP, 2018.
- **Lundberg Scott**, « A unified approach to explain the output of any machine learning model » [en ligne], disponible sur <https://github.com/slundberg/shap> [consulté le 8 septembre 2019].
- **Lundberg Scott M. et Su-In Lee**, « A unified approach to interpreting model predictions » dans I. Guyon (*et al.*), *Advances in Neural Information Processing Systems*, NIPS, 2017.
- **Molnar Christoph**, « Interpretable machine learning Shapley Value » [en ligne], disponible sur <https://christophm.github.io/interpretable-ml-book/shap.html> [consulté le 8 septembre 2019].
- **Monbet Valerie**, « Tests statistiques » [en ligne], Université de Rennes 1, 2009, disponible sur https://perso.univ-rennes1.fr/valerie.monbet/doc/cours/Cours_Tests_2009.pdf [consulté le 8 septembre 2019].

- **Rakotomalala Ricco**, « Régressions ridge et elasticnet sous R » [en ligne], Université Lumière Lyon 2, 2018, disponible sur http://eric.univ-lyon2.fr/ricco/tanagra/fichiers/fr_Tanagra_Ridge_Elasticnet_R.pdf [consulté le 8 septembre 2019].
- **Tibshirani Robert, Hastie Trevor et Wainwright Martin**, *Statistical Learning with Sparsity : The Lasso and Generalizations*, Chapman and Hall, 2015.
- **Wikistat**, « Agrégation de modèles » [en ligne], disponible sur <https://www.math.univ-toulouse.fr/besse/Wikistat/pdf/st-m-app-agreg.pdf> [consulté le 8 septembre 2019].

Annexe A

Intervalle de confiance sur les probabilités de souscription d'un produit

Dans la **partie 4.1.2**, des probabilités d'équipement vers un produit ont été calculées pour 14 classes d'individus. Ces probabilités correspondent à la moyenne au sein de la classe. Dans cette partie est calculé l'intervalle de confiance autour de ces moyennes afin de vérifier leur pertinence et leur stabilité.

L'estimateur de la moyenne converge en loi vers une loi normale, il est alors possible de calculer un intervalle de confiance asymptotique.

Soit $\hat{\theta}$ l'estimateur de la probabilité correspondant à la moyenne : $\sqrt{n} \times \frac{\hat{\theta} - \theta}{\theta \times (1 - \theta)} \rightarrow \mathcal{N}(0, 1)$

Et par le théorème de Slutsky on a : $\sqrt{n} \times \frac{\hat{\theta} - \theta}{\hat{\theta} \times (1 - \hat{\theta})} \rightarrow \mathcal{N}(0, 1)$

Ainsi l'intervalle de confiance asymptotique vaut : $\left[\hat{\theta} \pm q_{1-\frac{\alpha}{2}} \times \frac{\hat{\theta} \times (1 - \hat{\theta})}{\sqrt{n}} \right]$

Où $q_{1-\frac{\alpha}{2}}$ est le quantile de la loi normale d'ordre $q_{1-\frac{\alpha}{2}}$

Le **tableau A.1** affiche l'erreur de la prédiction pour un niveau de confiance à 5 % (i.e. : $q_{1-\frac{\alpha}{2}} = 1.96$).

La probabilité a une erreur en général inférieure à 1 %. L'erreur est supérieure à 1 % pour les classes présentant peu d'individus. Toutefois, l'erreur maximale est de 4 % pour la classe *PRO Vie*.

		Erreur de la probabilité pour un niveau de confiance de 95 %							
Classe du client	Nombre d'individus dans la classe	Automobile	Autre IARD	Engins Spéciaux	Epargne	Retraite	MRC	MRH & PNO	Prévoyance
PE-ENTIARD00	5351	0%	1%	0%	0%	0%	0%	0%	0%
PE-ENTIARD01	2865	0%	1%	0%	0%	0%	0%	0%	0%
PE-ENTIARD10	5444	1%	0%	0%	0%	0%	0%	0%	0%
PE-ENTIARD11	1320	1%	1%	1%	0%	0%	0%	0%	0%
PE-ENTMixte	1241	1%	1%	0%	0%	1%	1%	0%	0%
PE-ENTPers	956	1%	1%	0%	1%	1%	1%	0%	2%
PE-ENTVie	263	0%	1%	0%	2%	1%	1%	0%	1%
PROIARD00	2295	1%	0%	0%	0%	0%	0%	1%	1%
PROIARD01	661	2%	1%	1%	0%	0%	0%	1%	1%
PROIARD10	4948	1%	0%	0%	0%	0%	0%	0%	0%
PROIARD11	566	2%	1%	1%	0%	0%	0%	1%	0%
PROMixte	1392	1%	0%	0%	0%	0%	0%	1%	1%
PROPers	1203	1%	0%	0%	0%	0%	0%	1%	1%
PROVie	164	1%	0%	0%	3%	0%	0%	0%	4%

Figure A.1 – Erreur de la prédiction de la probabilité pour un seuil de confiance à 95 %

En conclusion, l'erreur associée aux probabilités est considérée comme faible.

Annexe B

Graphiques d'ajustement des lois de probabilité pour chaque feuille des arbres de régression

Cette annexe présente les graphiques et statistiques d'ajustement pour chacune des feuilles des arbres de décision de la **partie 4.3**.

B.1 Feuilles de l'arbre pour les produits *MRH* & *PNO*

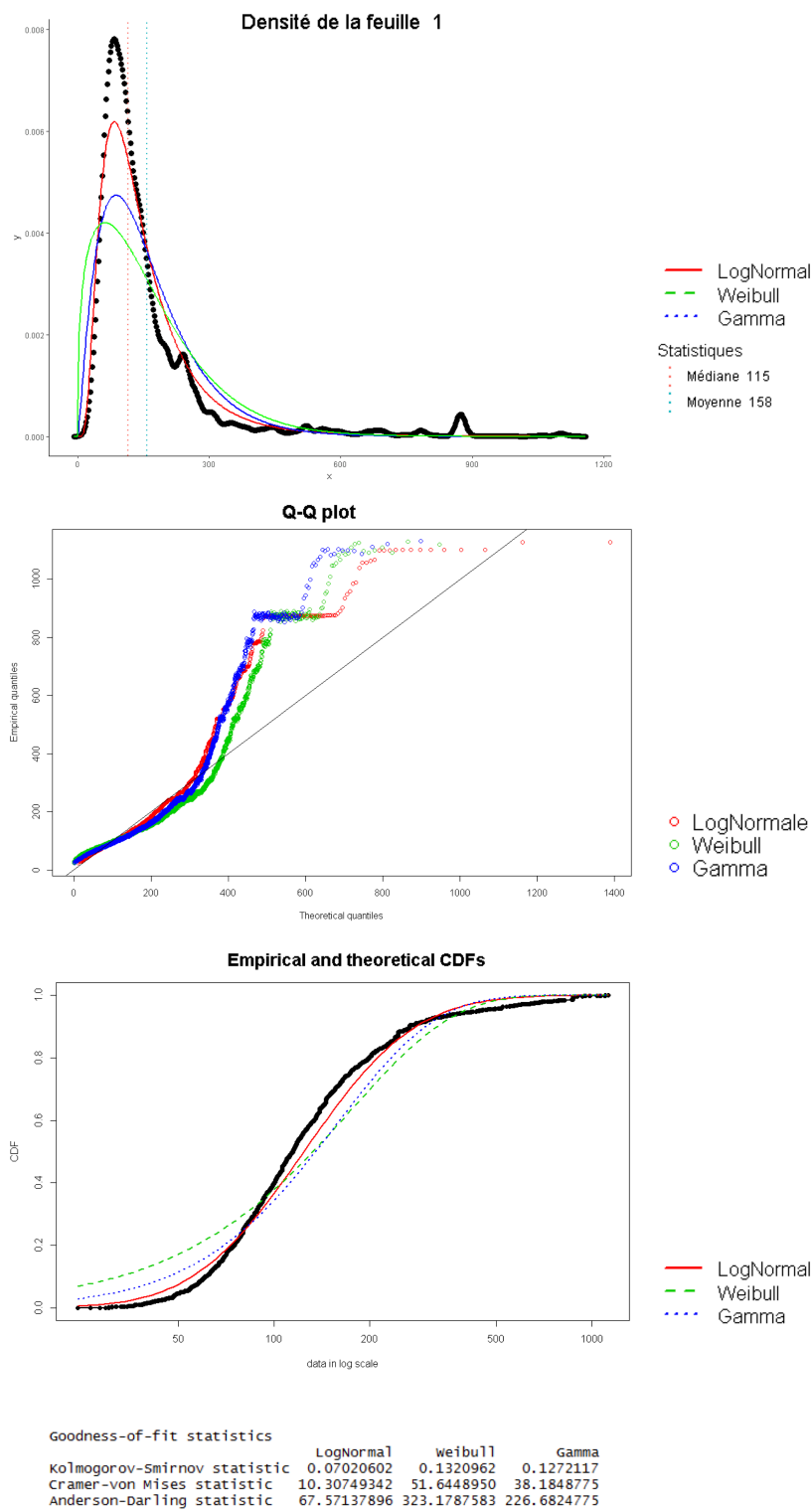


Figure B.1 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 1

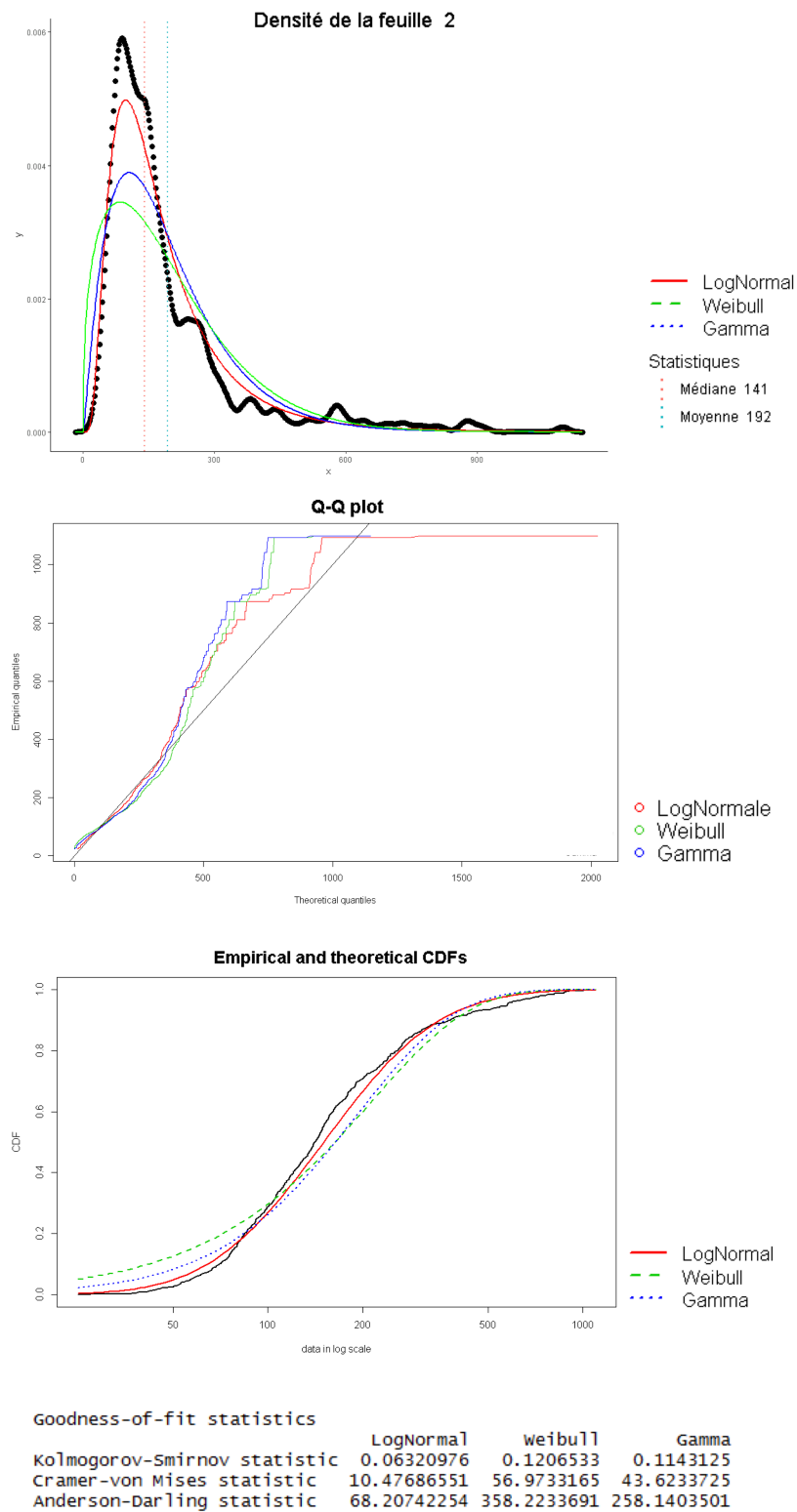


Figure B.2 – Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 2

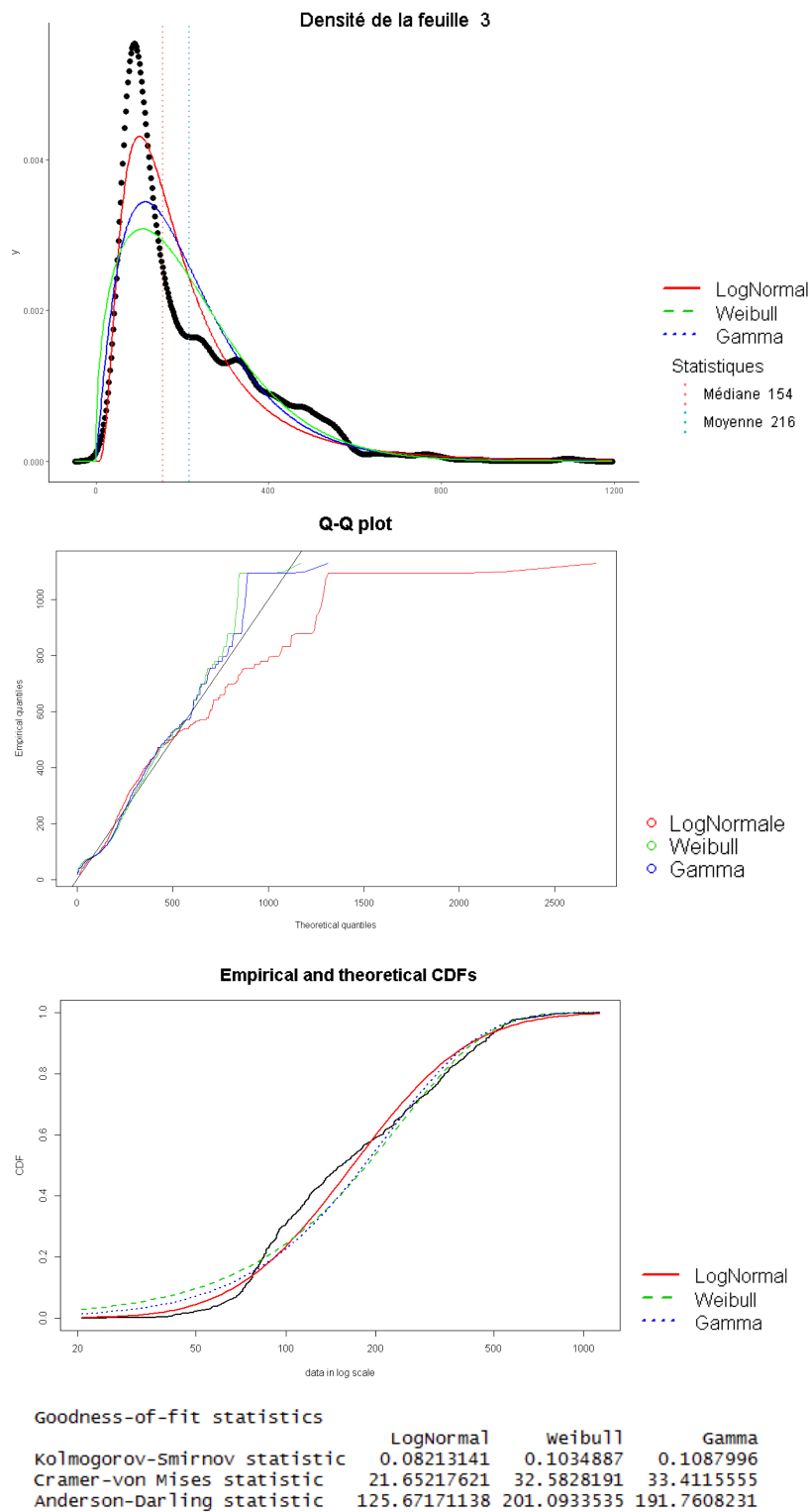


Figure B.3 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 3

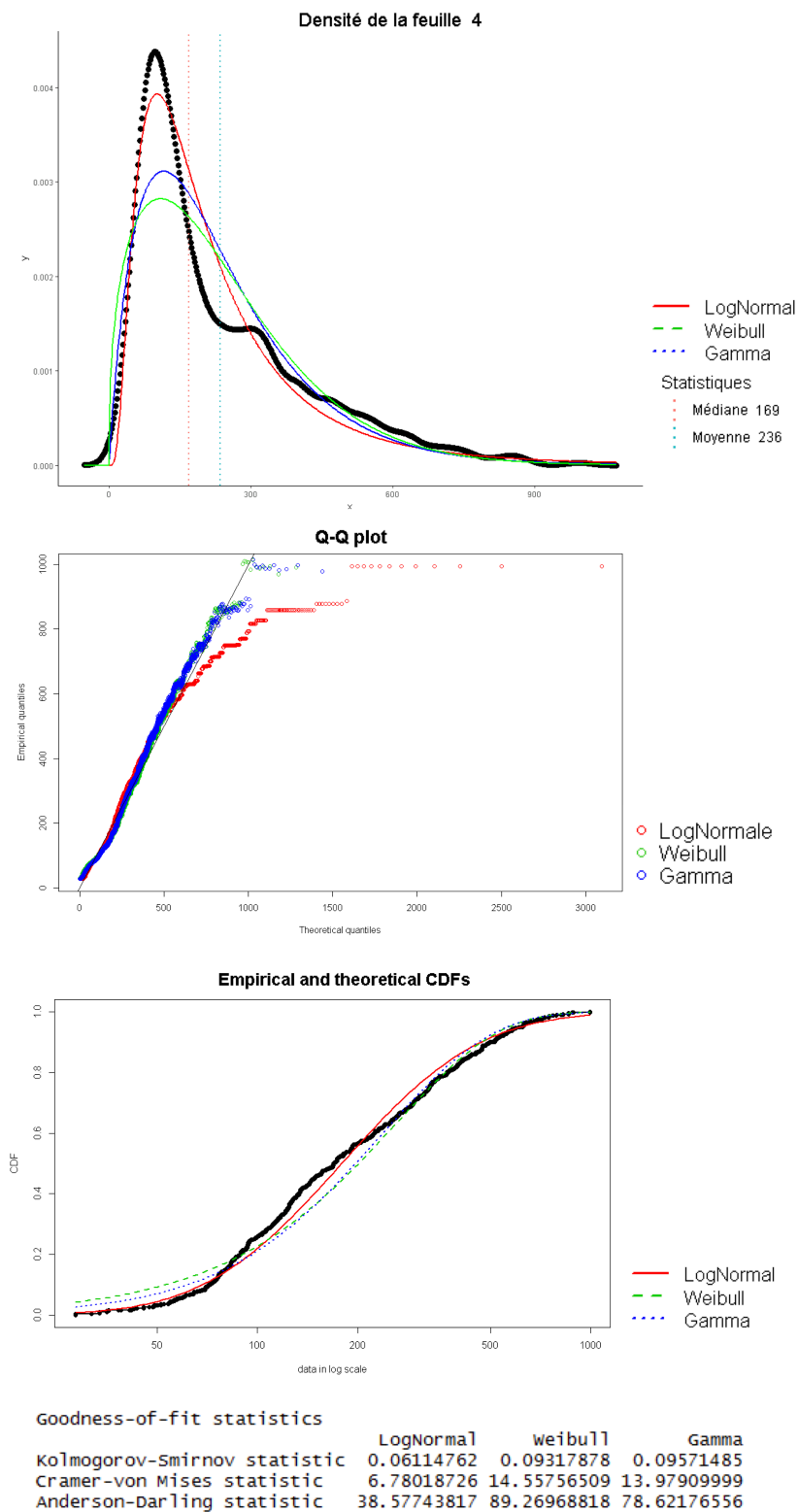


Figure B.4 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 4

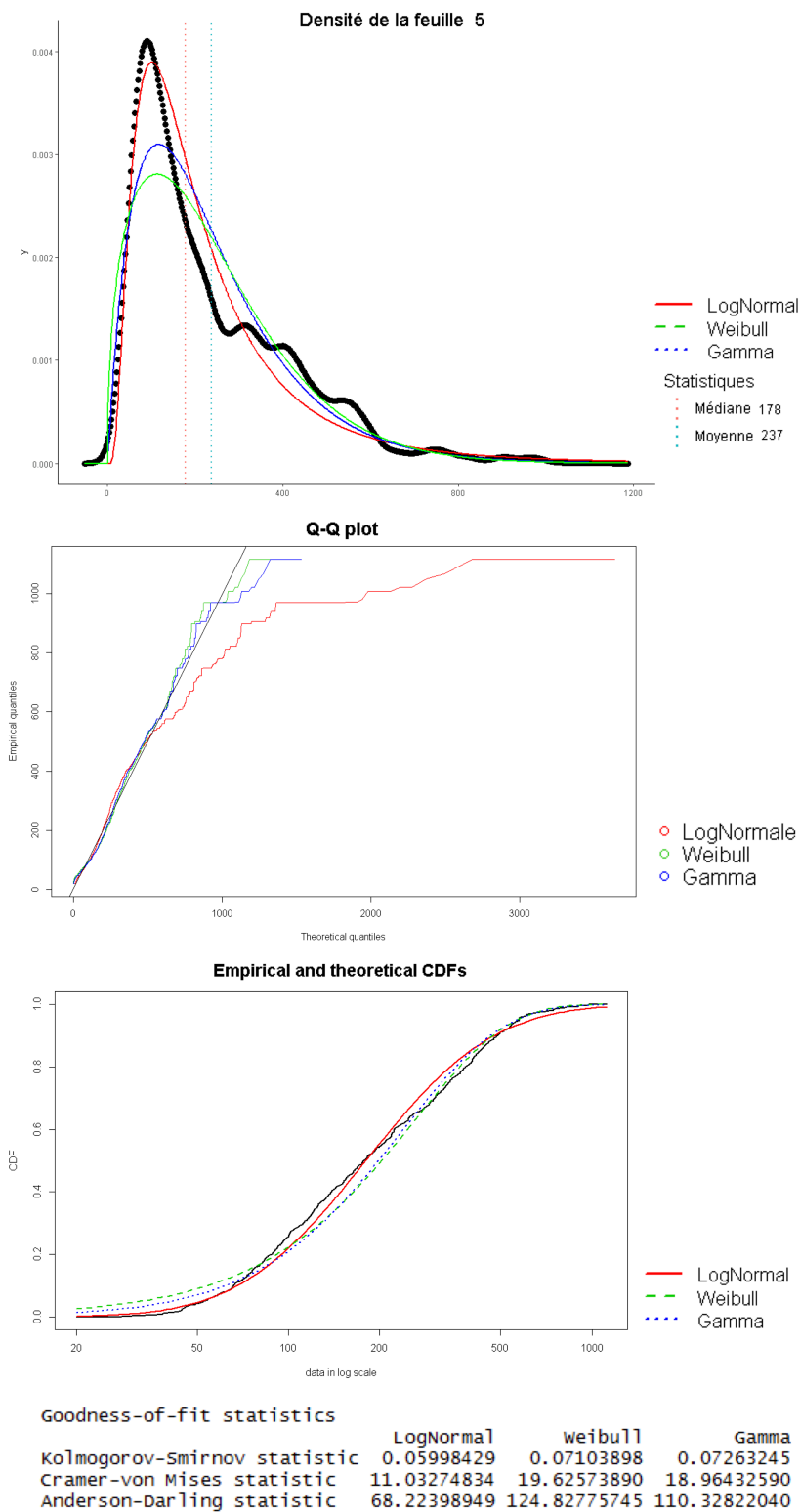


Figure B.5 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 5

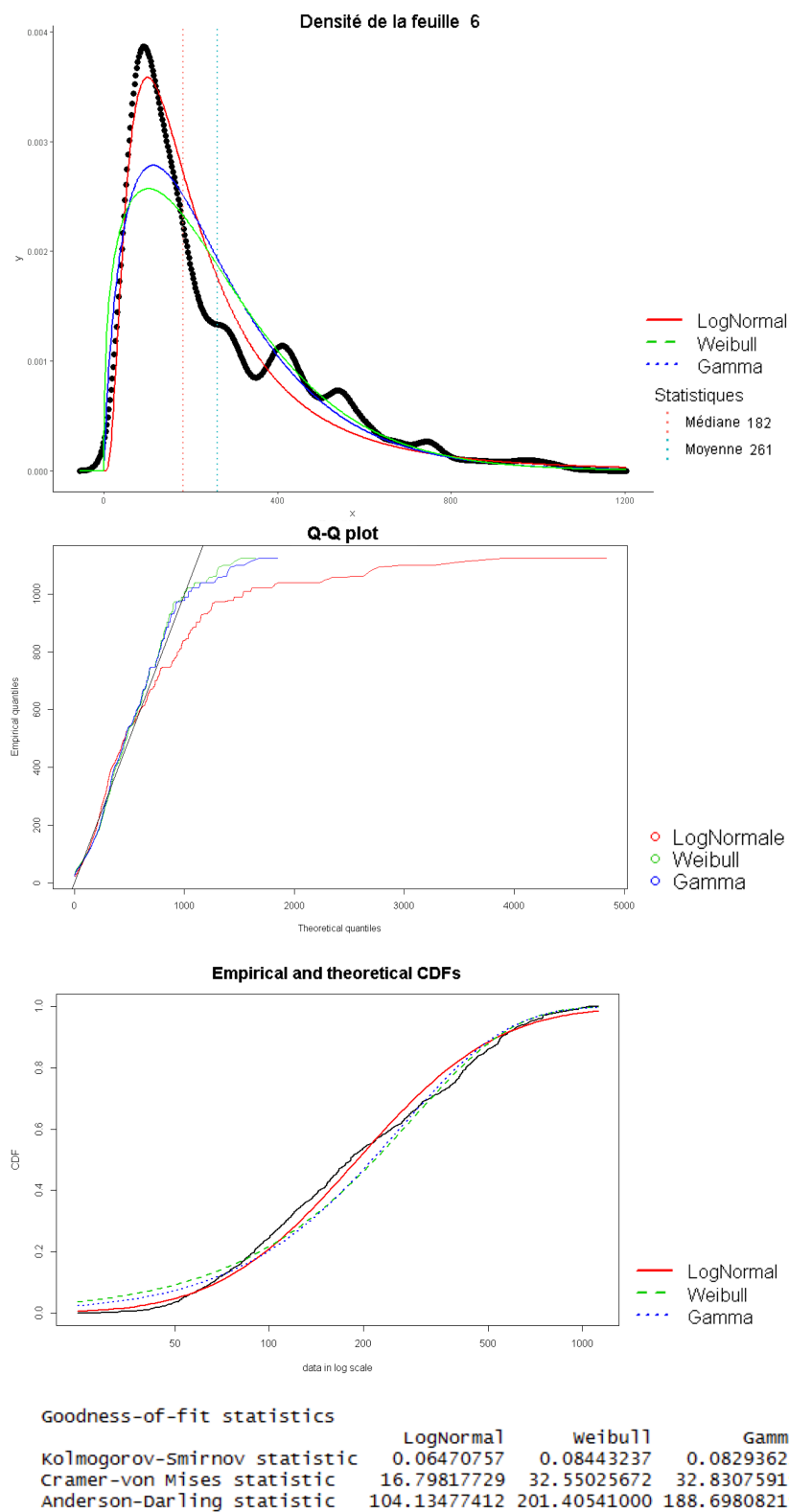


Figure B.6 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 6

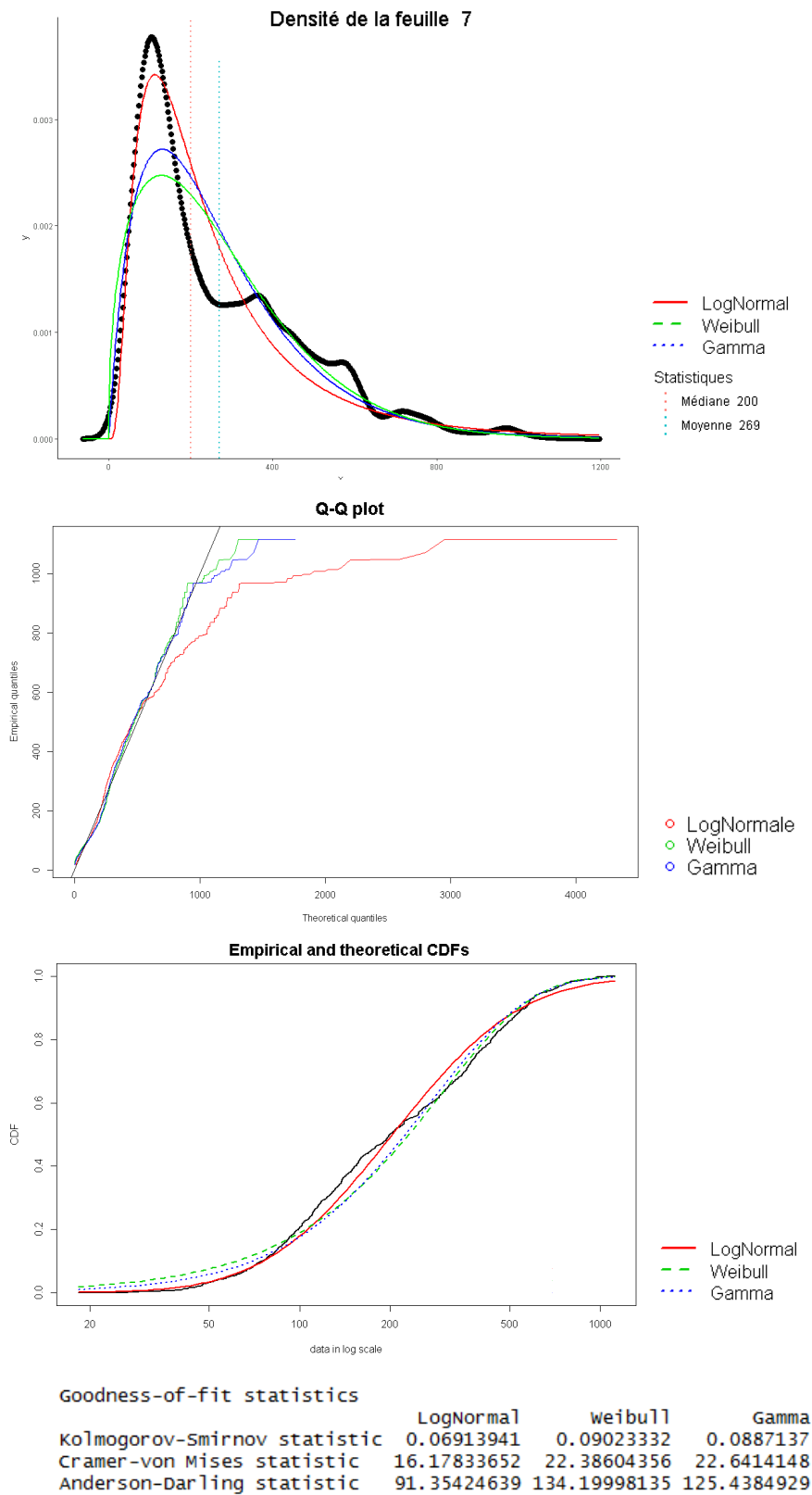


Figure B.7 – Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 7

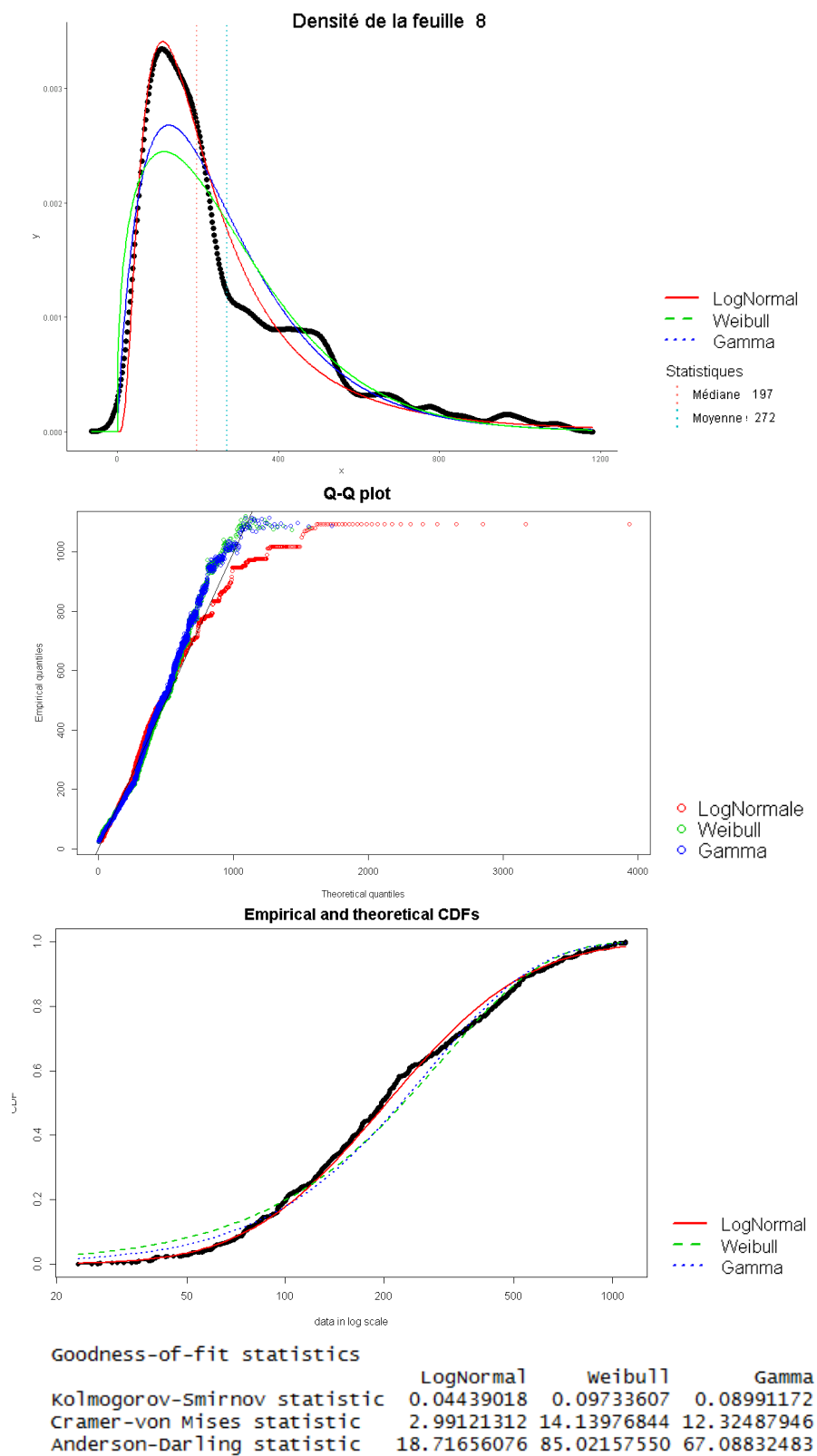


Figure B.8 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 8

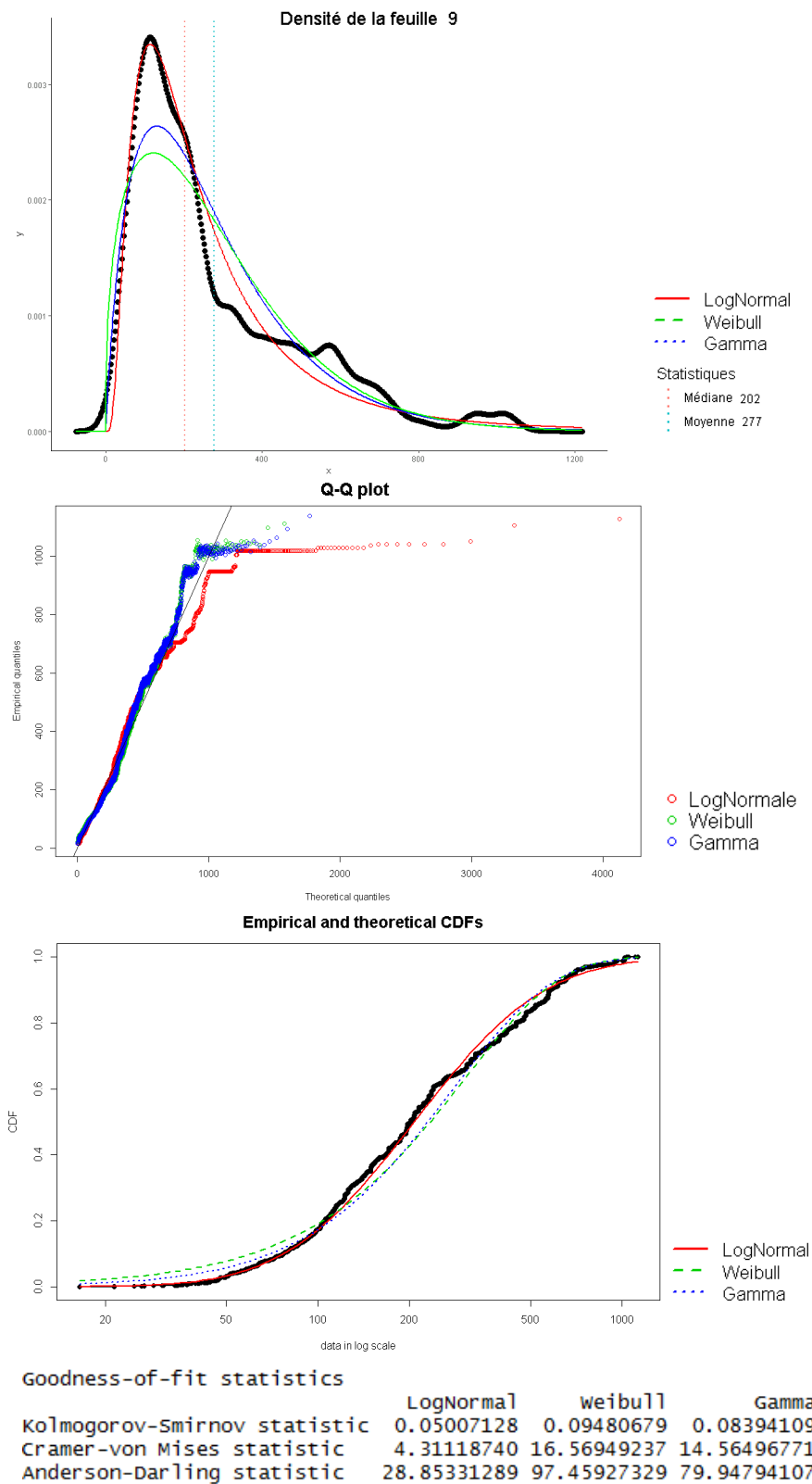


Figure B.9 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 9

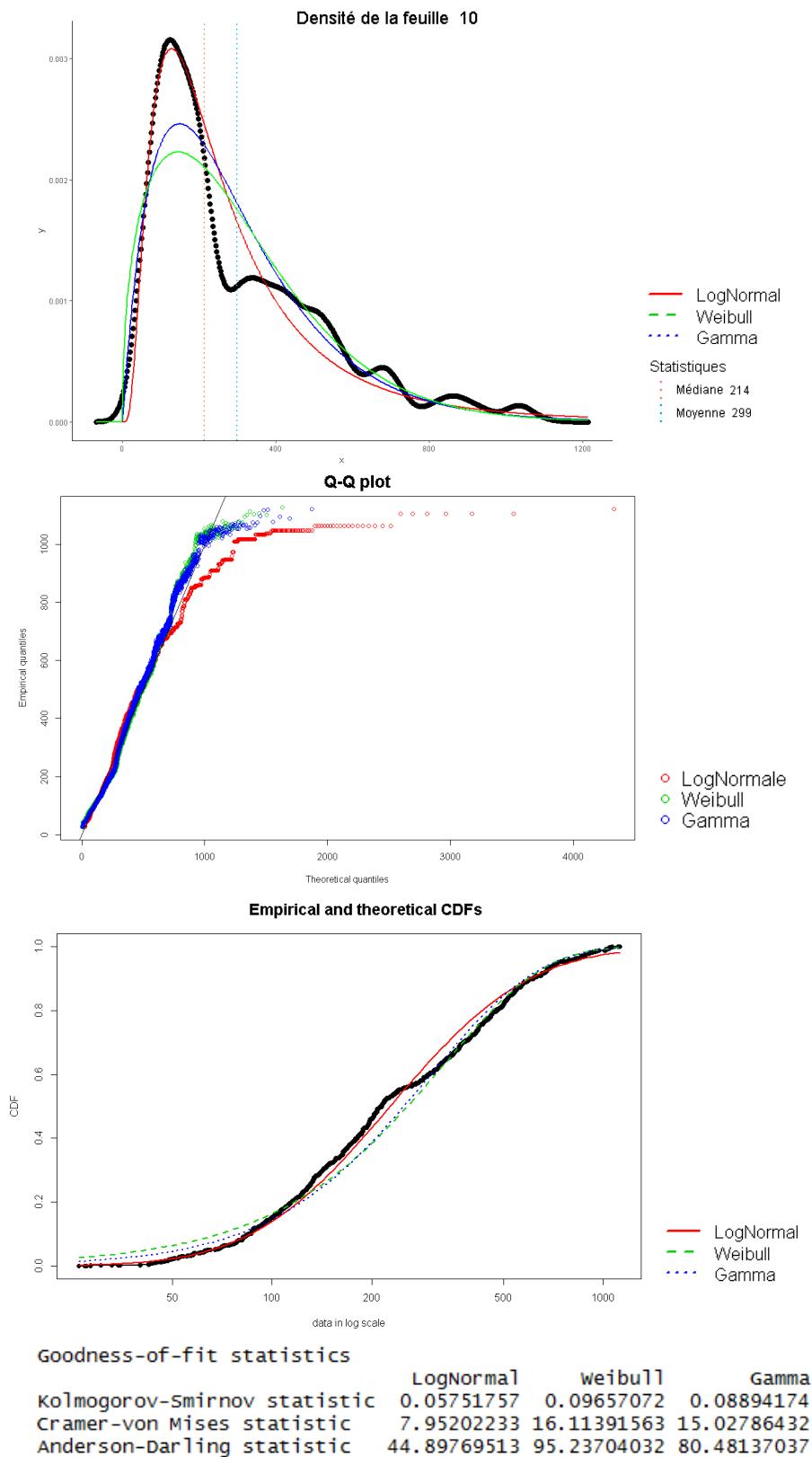


Figure B.10 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 10

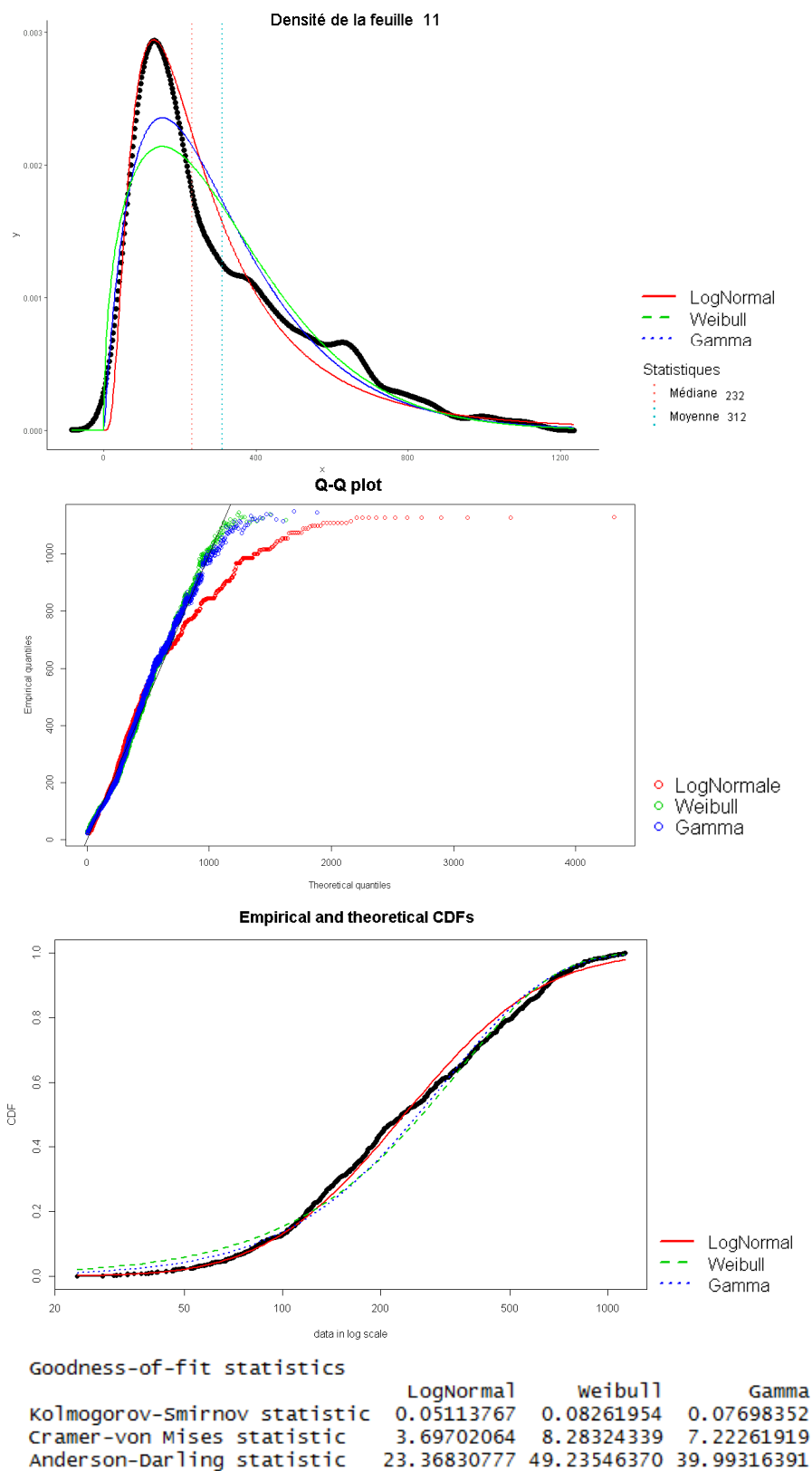


Figure B.11 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 11

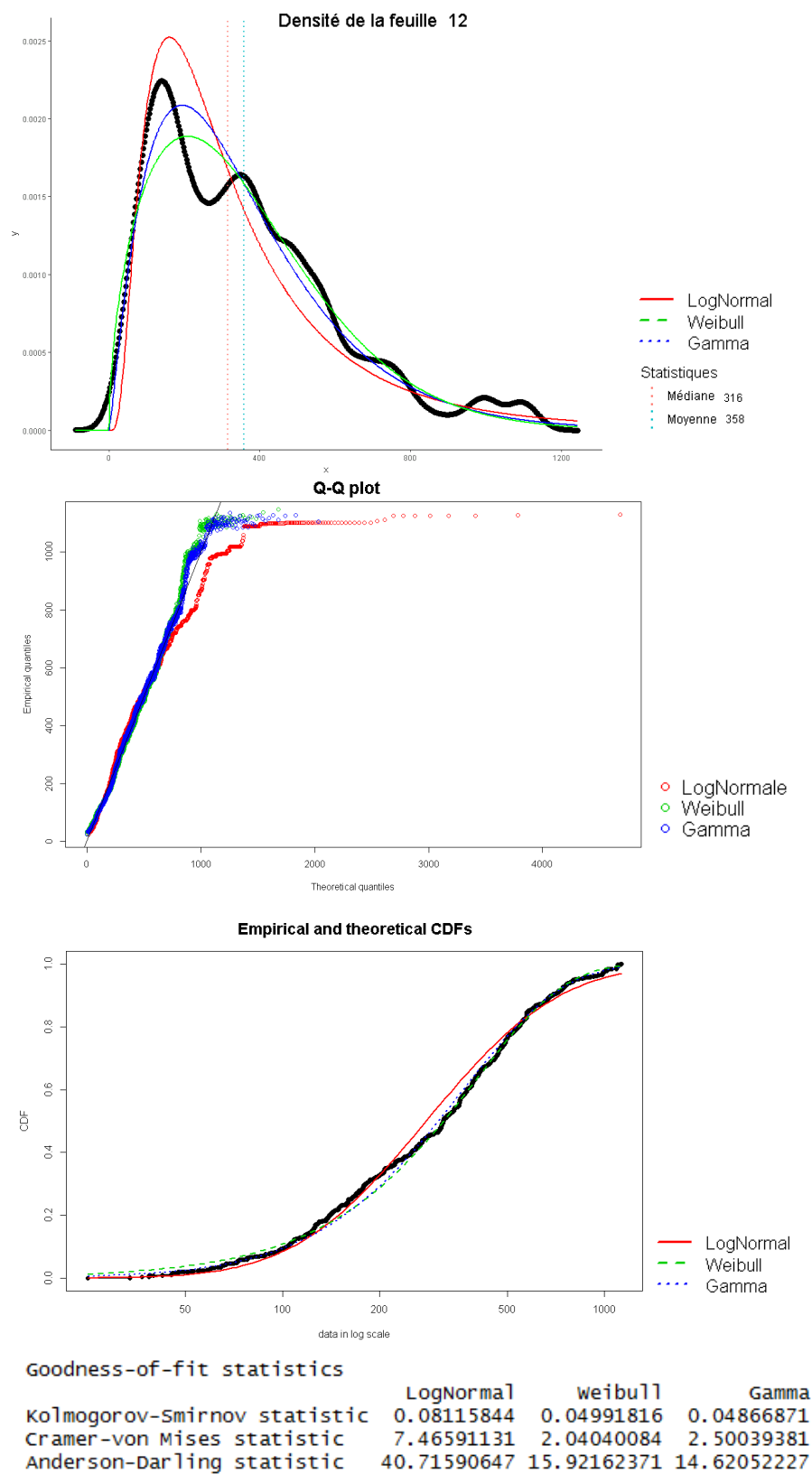


Figure B.12 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 12

B.2 Feuilles de l'arbre pour les produits *Auto-mobile hors Engins Spéciaux*

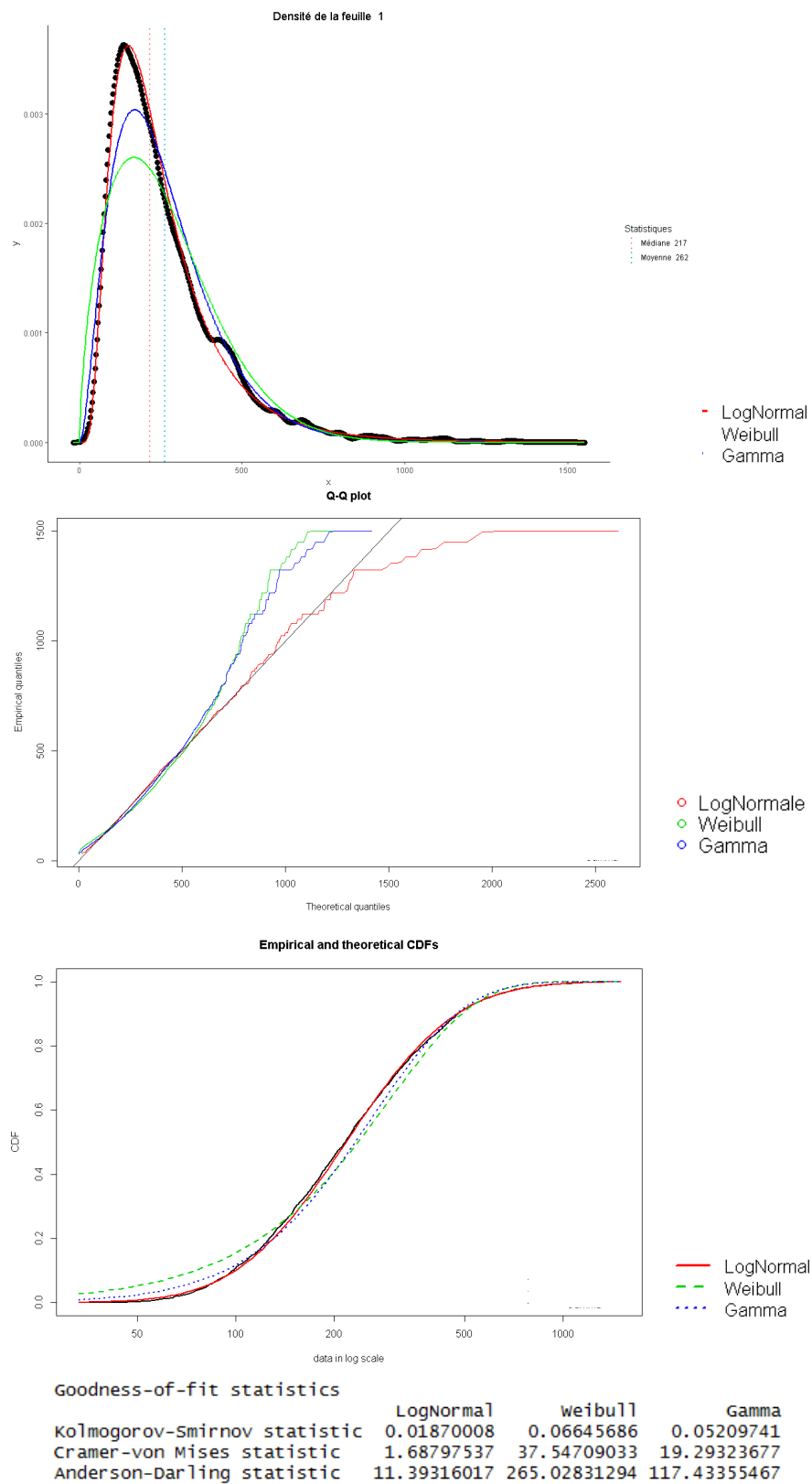


Figure B.13 – Graphiques et statistiques d'ajustement des lois de probabilités de la
Feuille 1

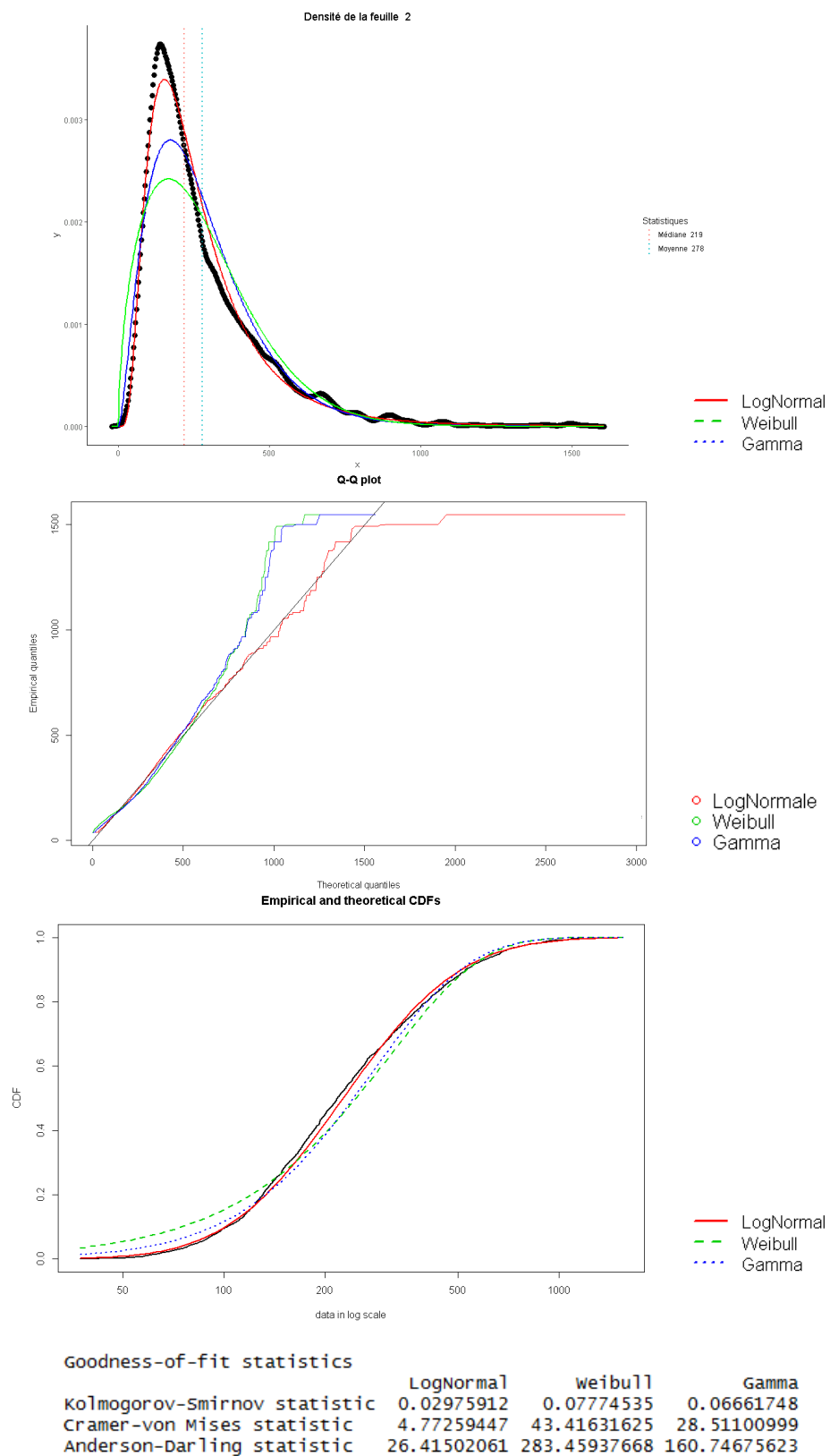


Figure B.14 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 2

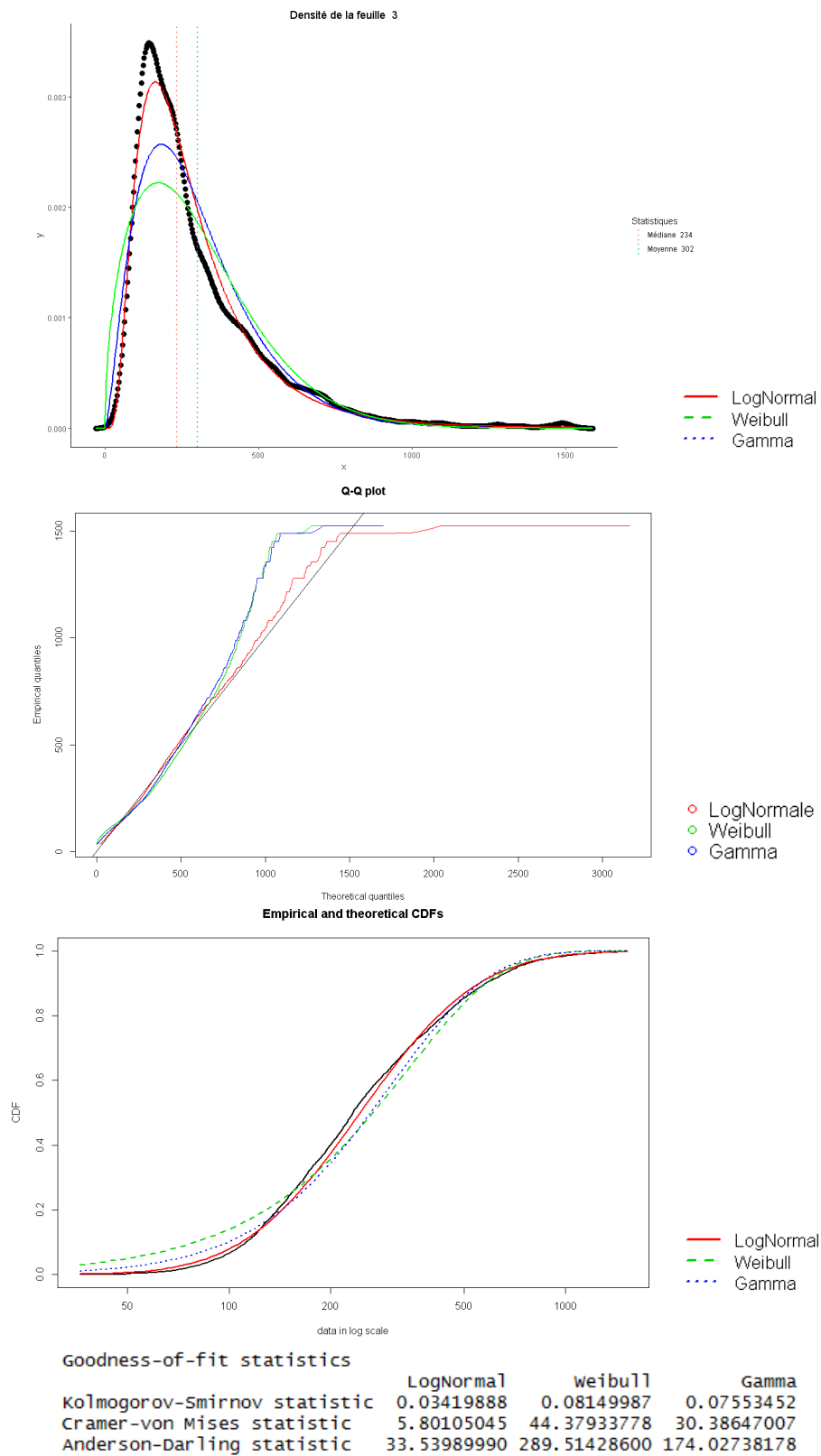


Figure B.15 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 3

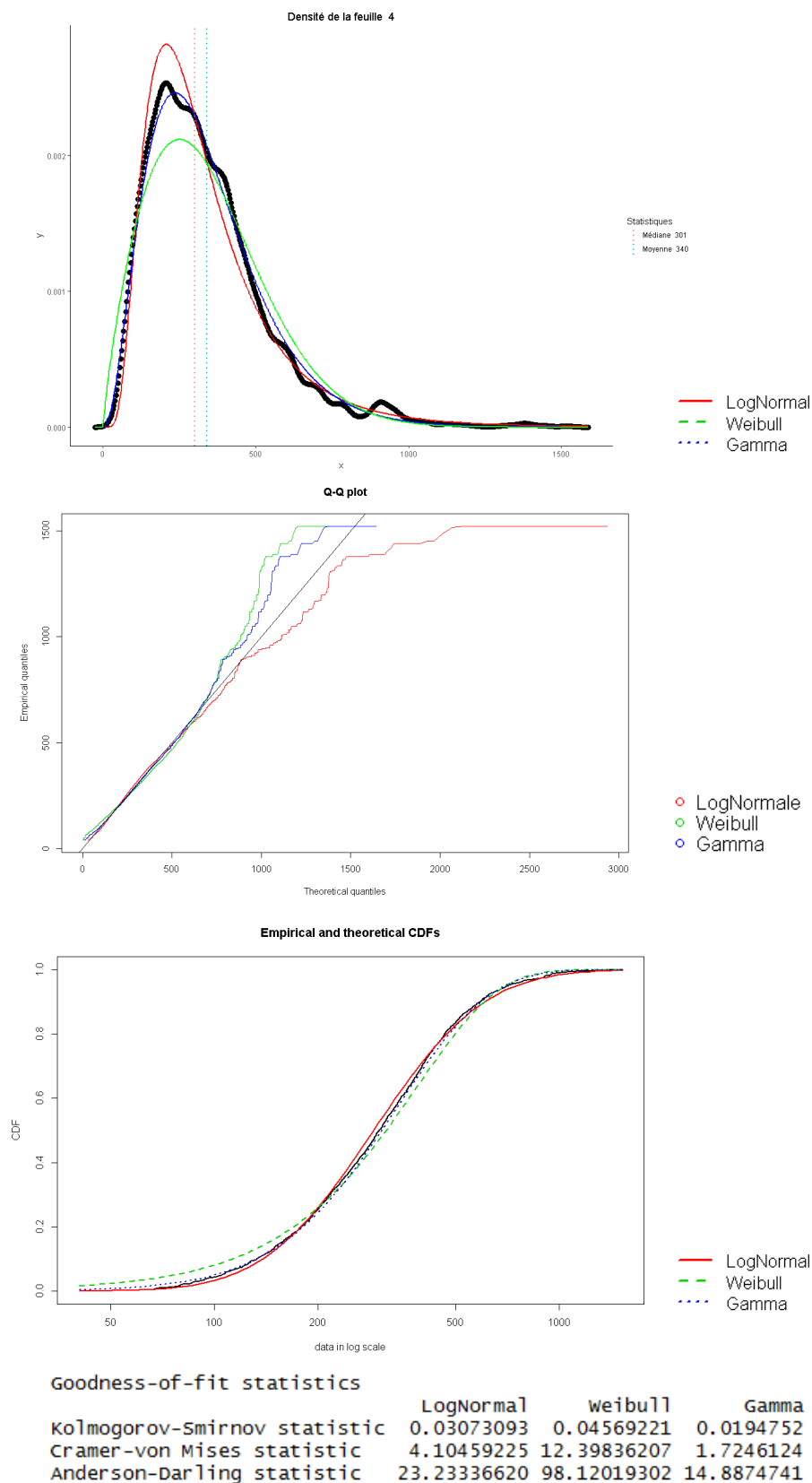


Figure B.16 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 4

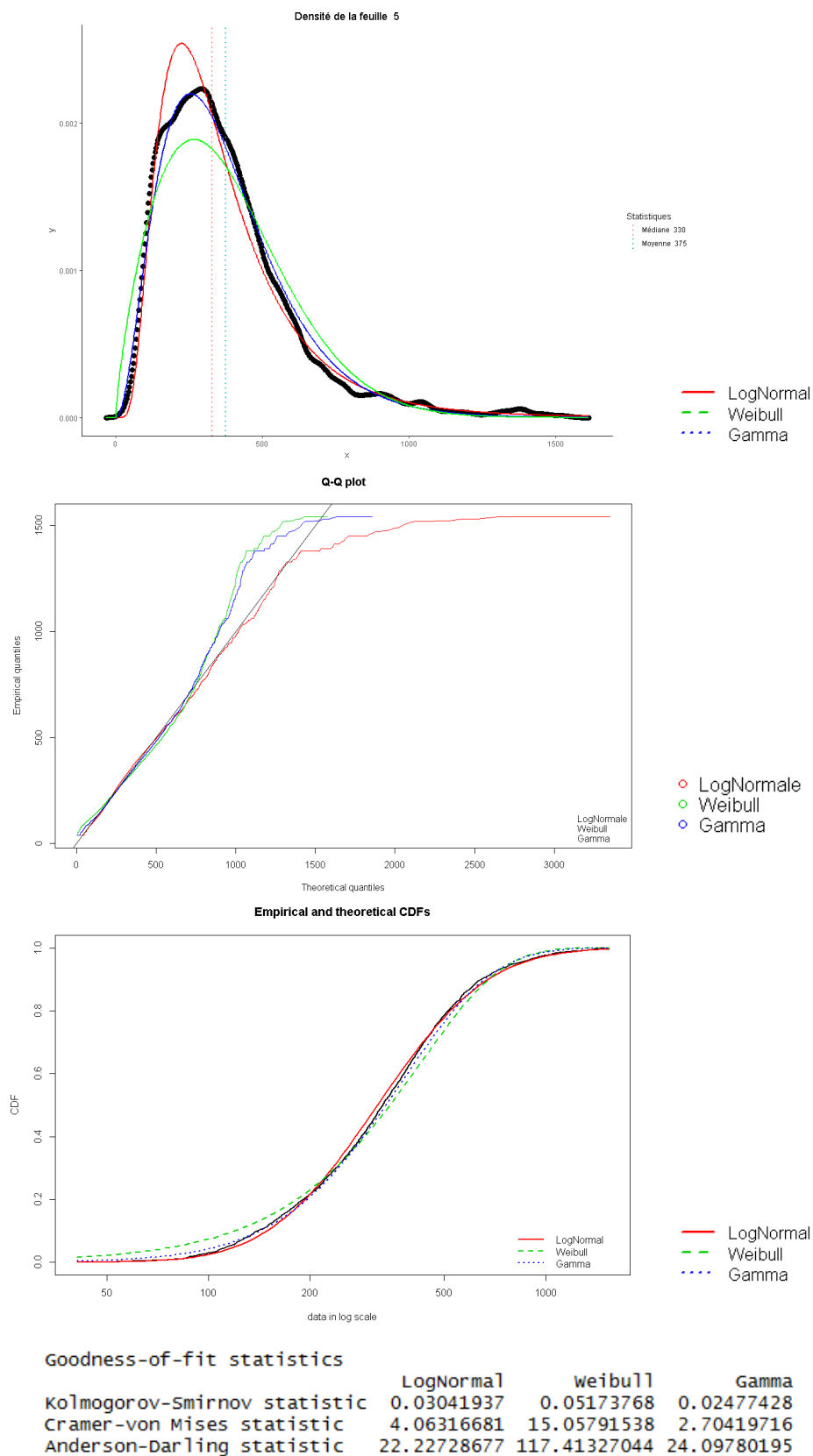


Figure B.17 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 5

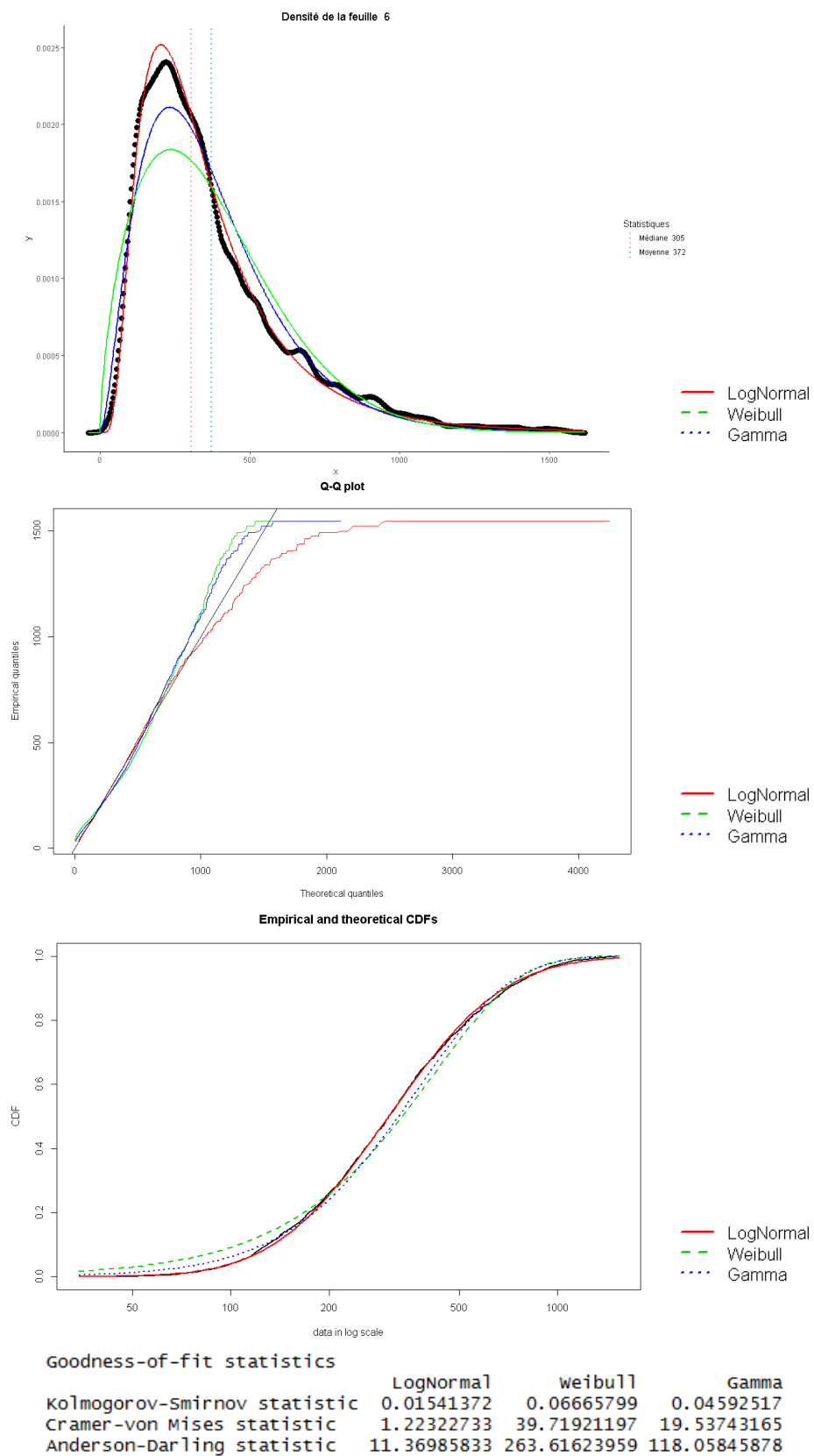


Figure B.18 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 6

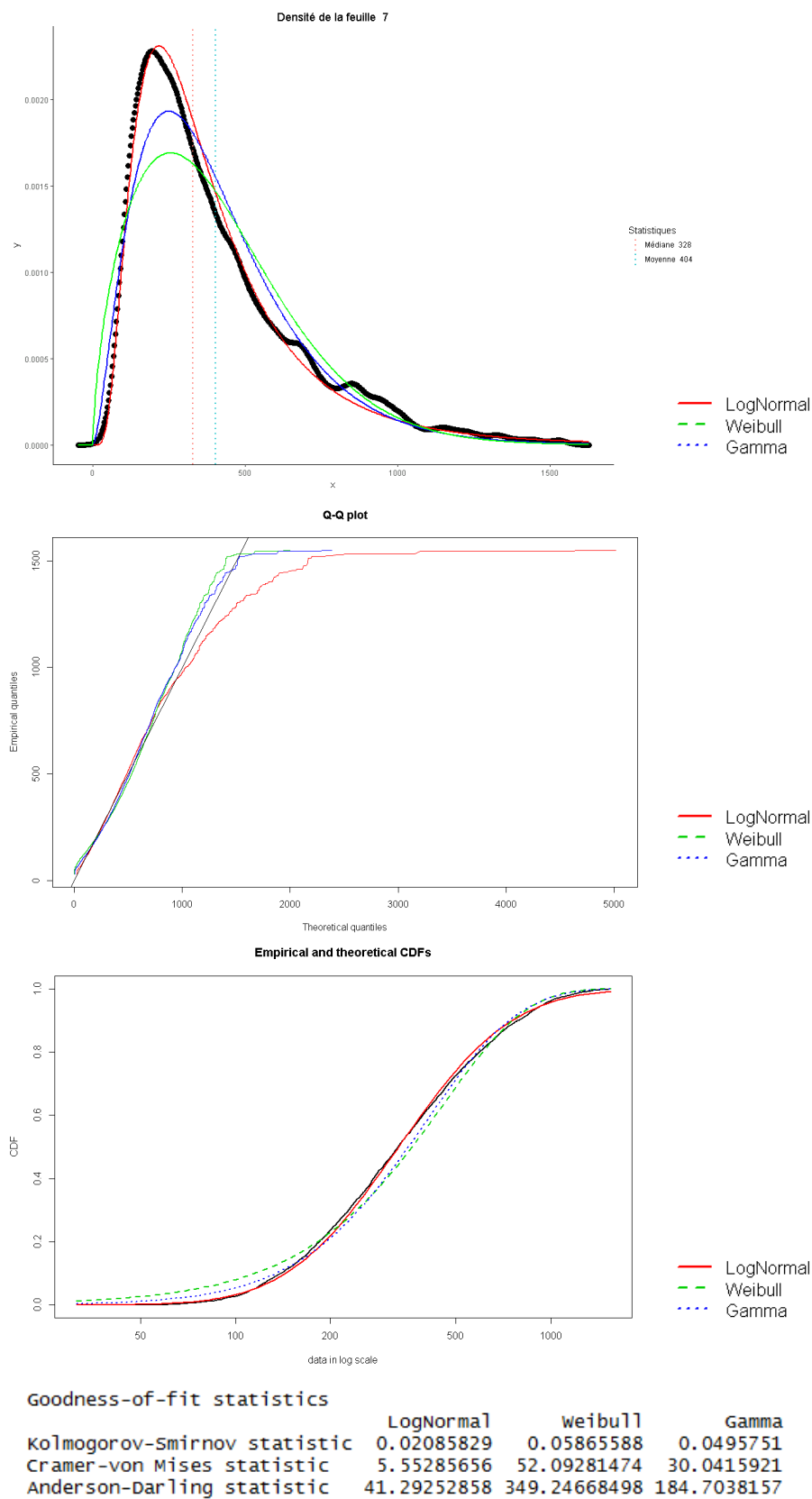


Figure B.19 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 7

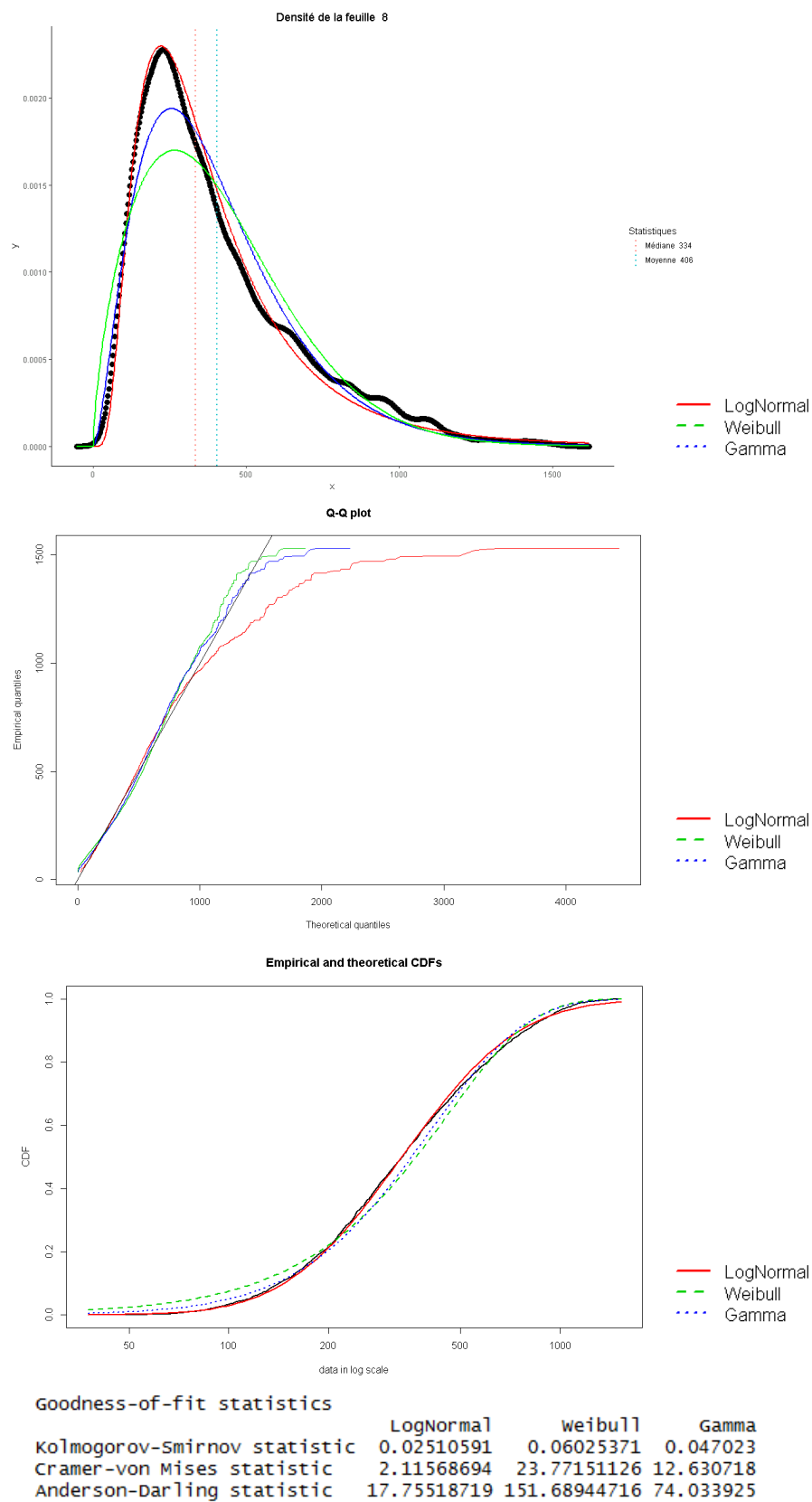


Figure B.20 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 8

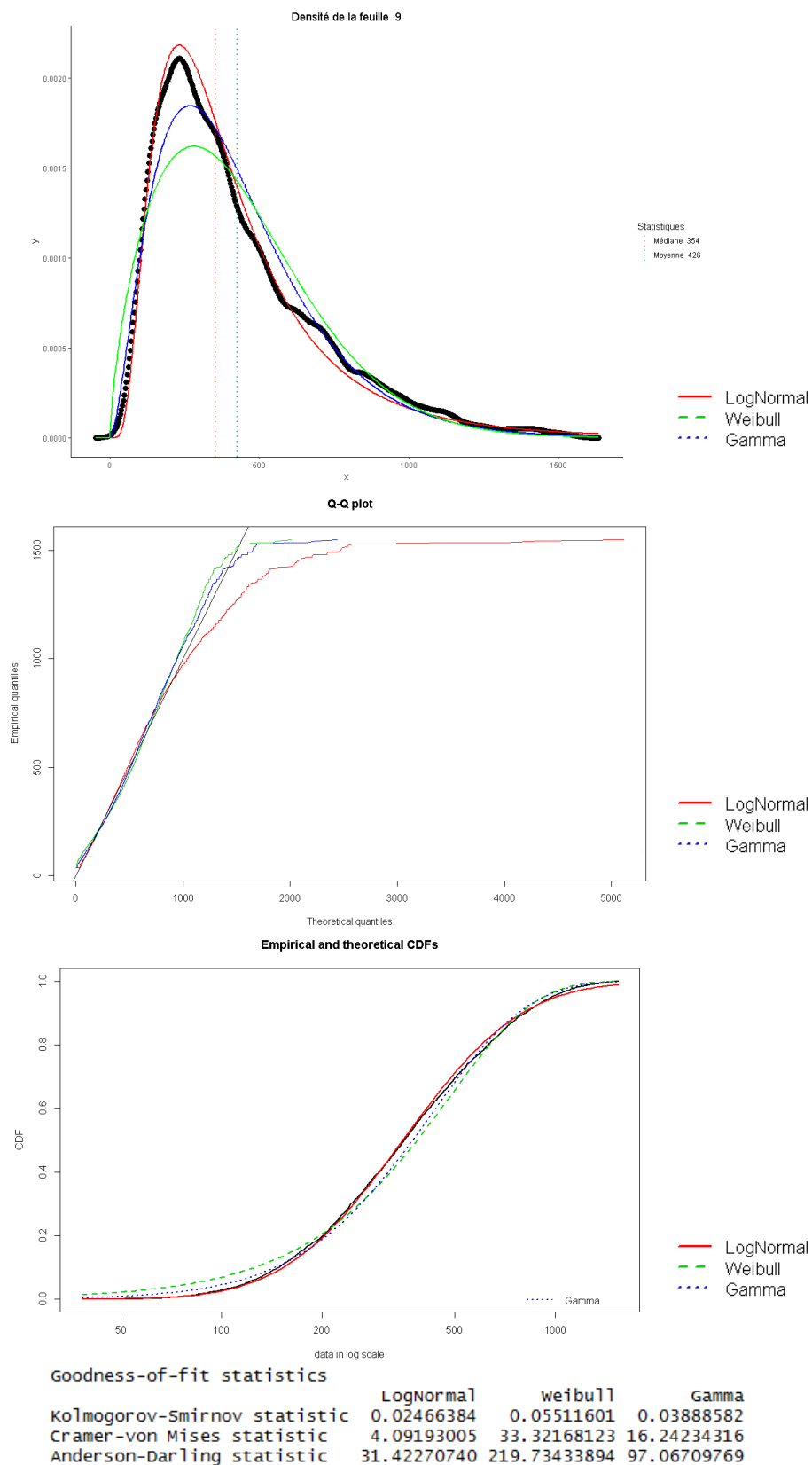


Figure B.21 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 9

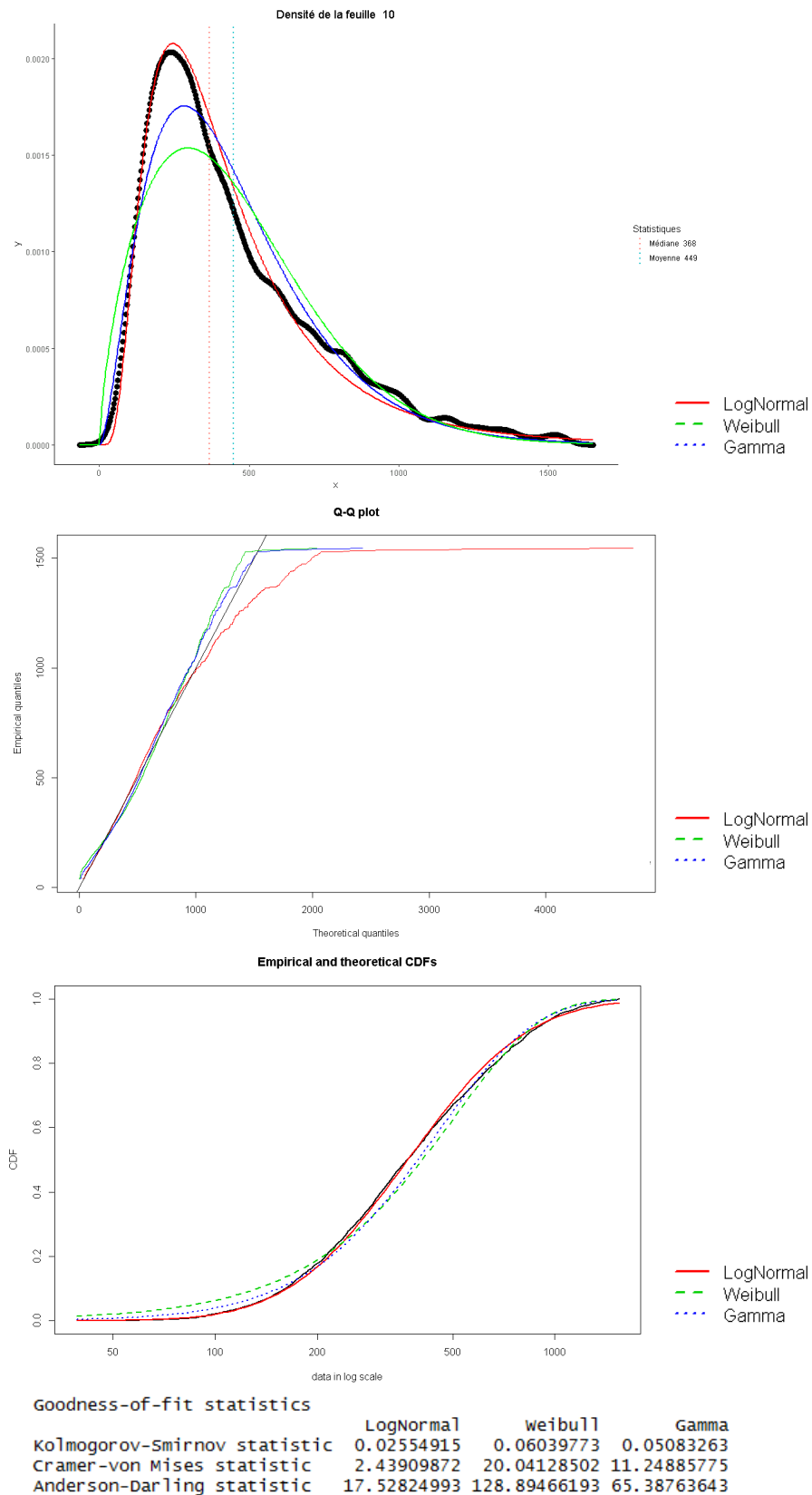


Figure B.22 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 10

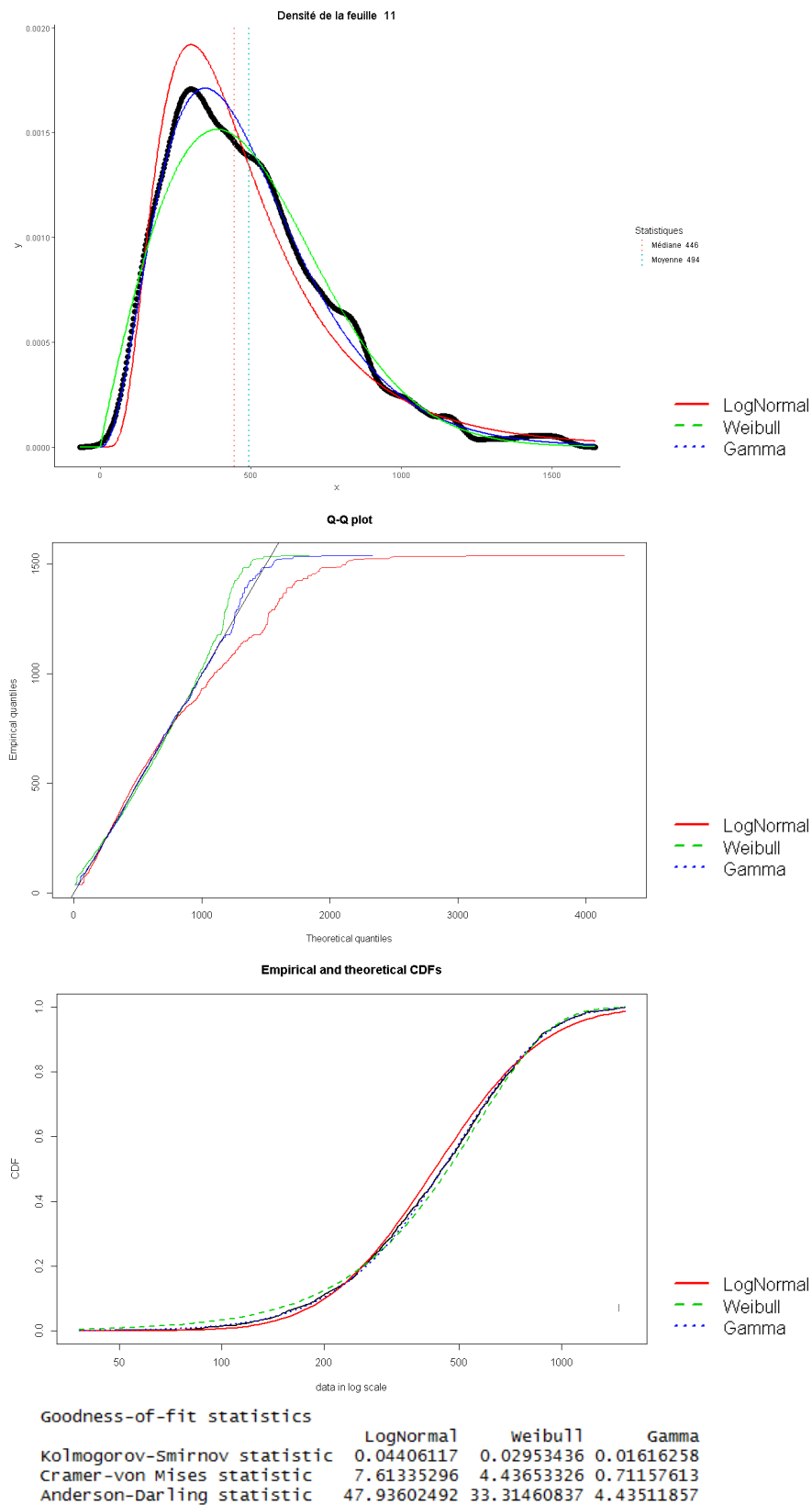


Figure B.23 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 11

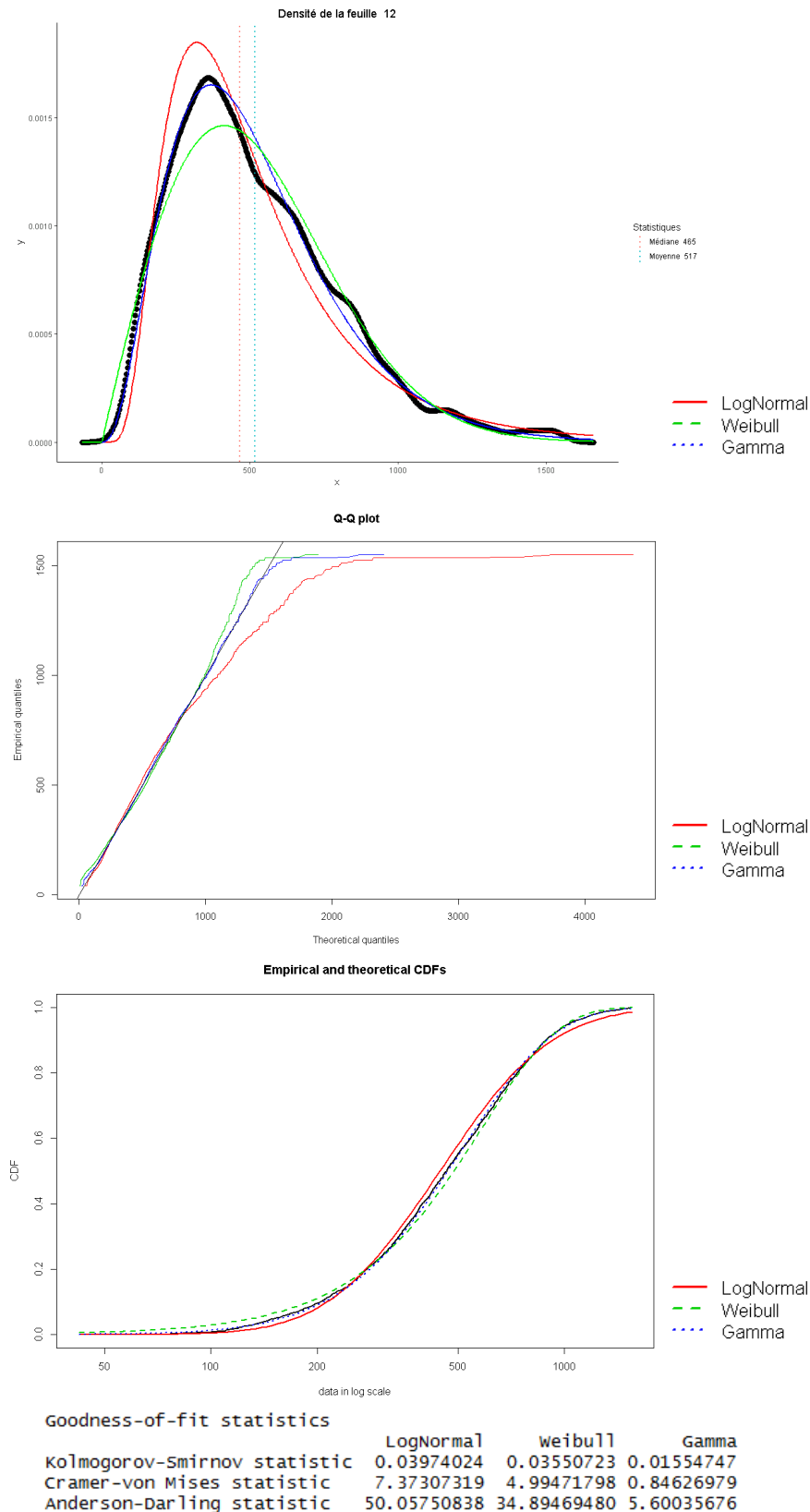


Figure B.24 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 12

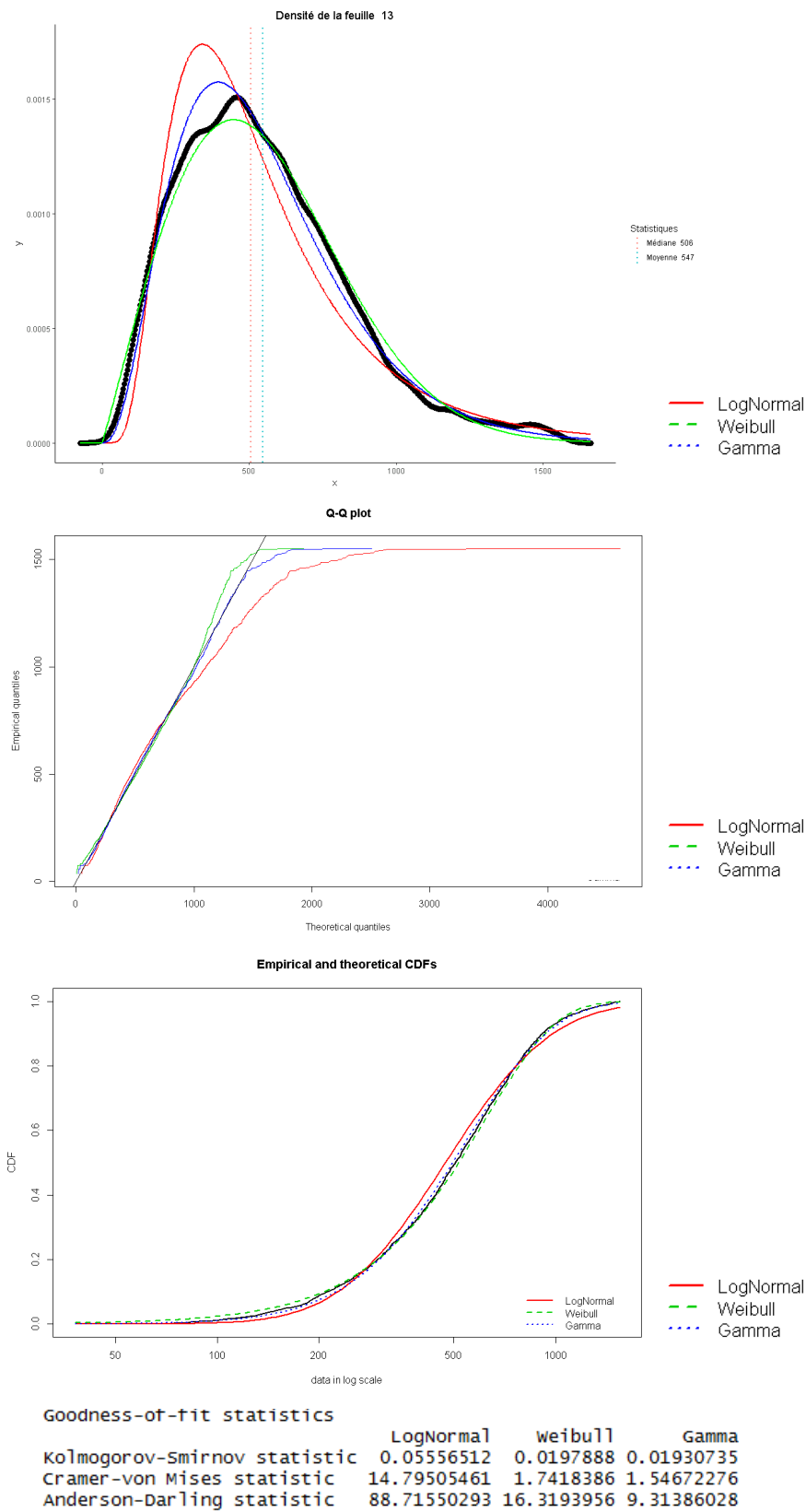


Figure B.25 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 13

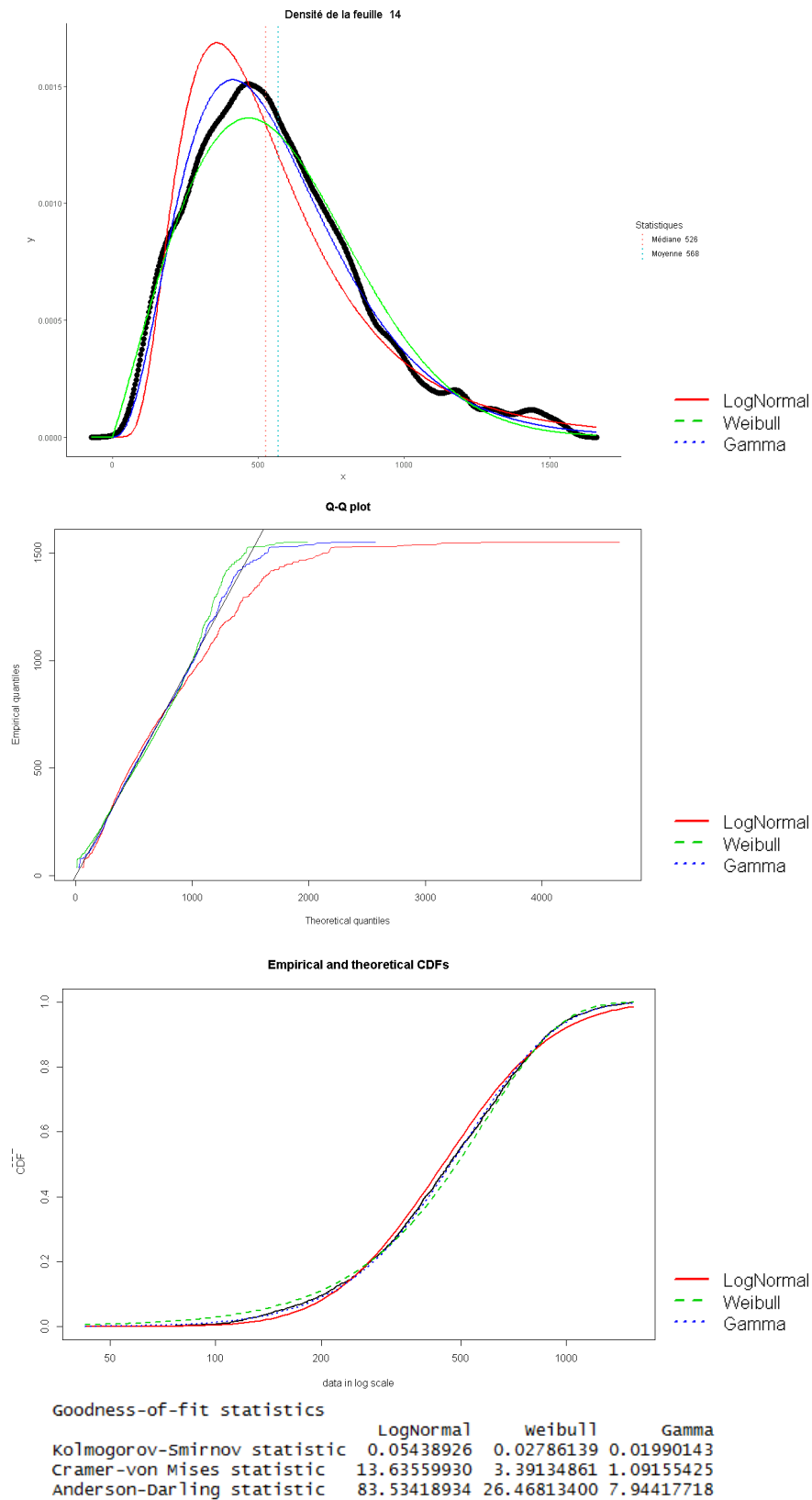


Figure B.26 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 14

B.3 Feuilles de l'arbre pour les produits *Engins Spéciaux*

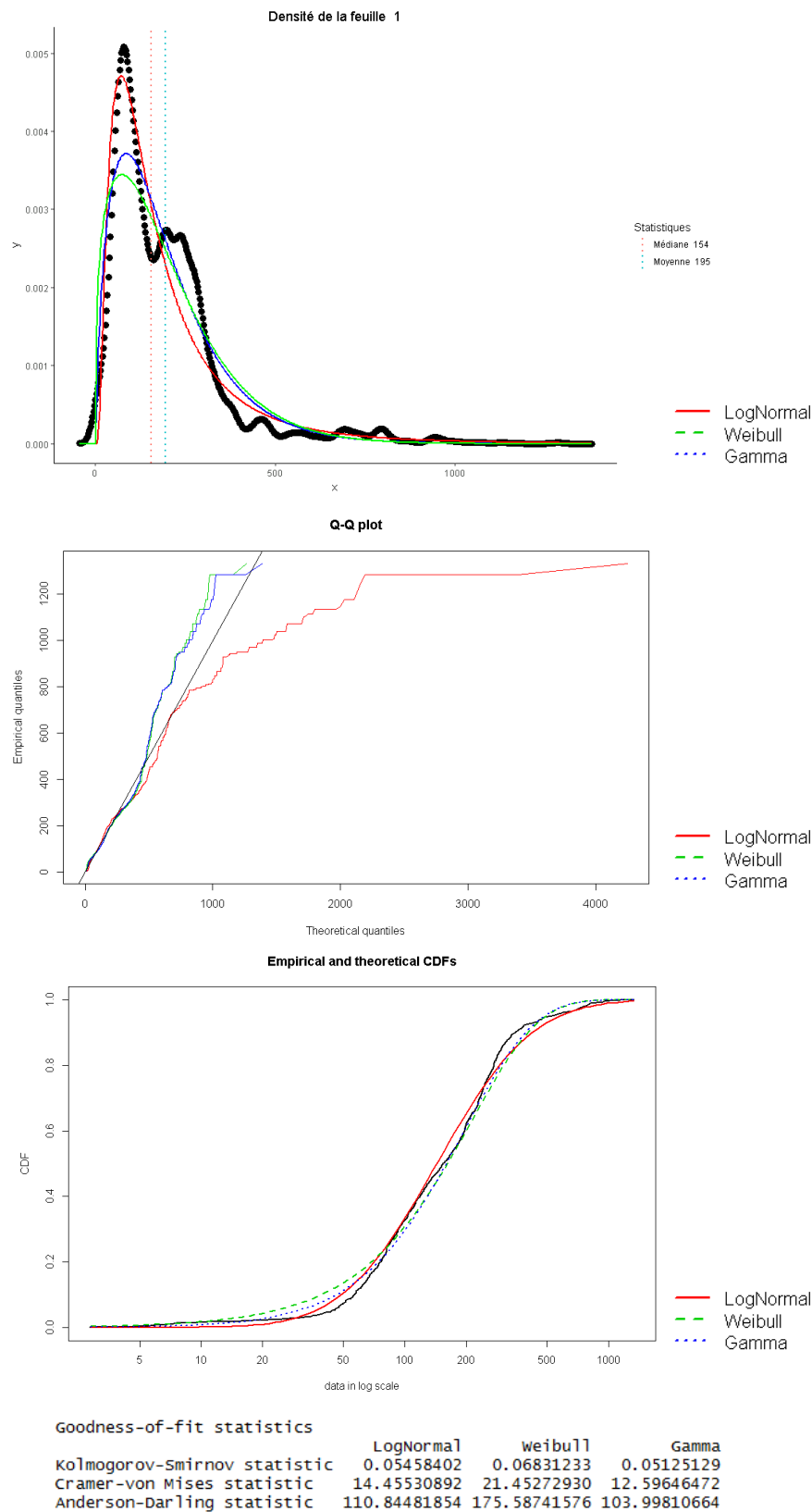


Figure B.27 – Graphiques et statistiques d'ajustement des lois de probabilités de la
Feuille 1

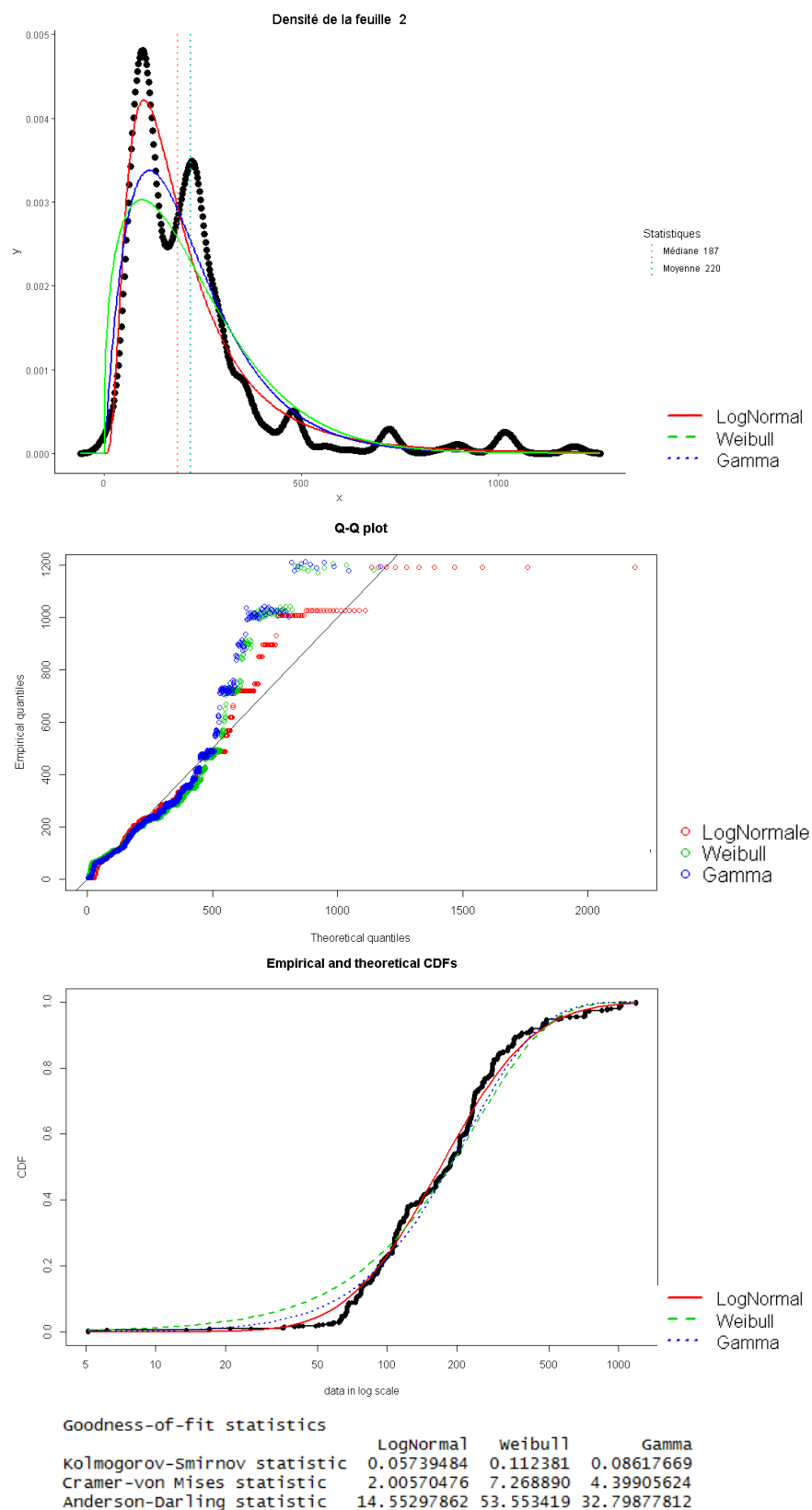


Figure B.28 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 2

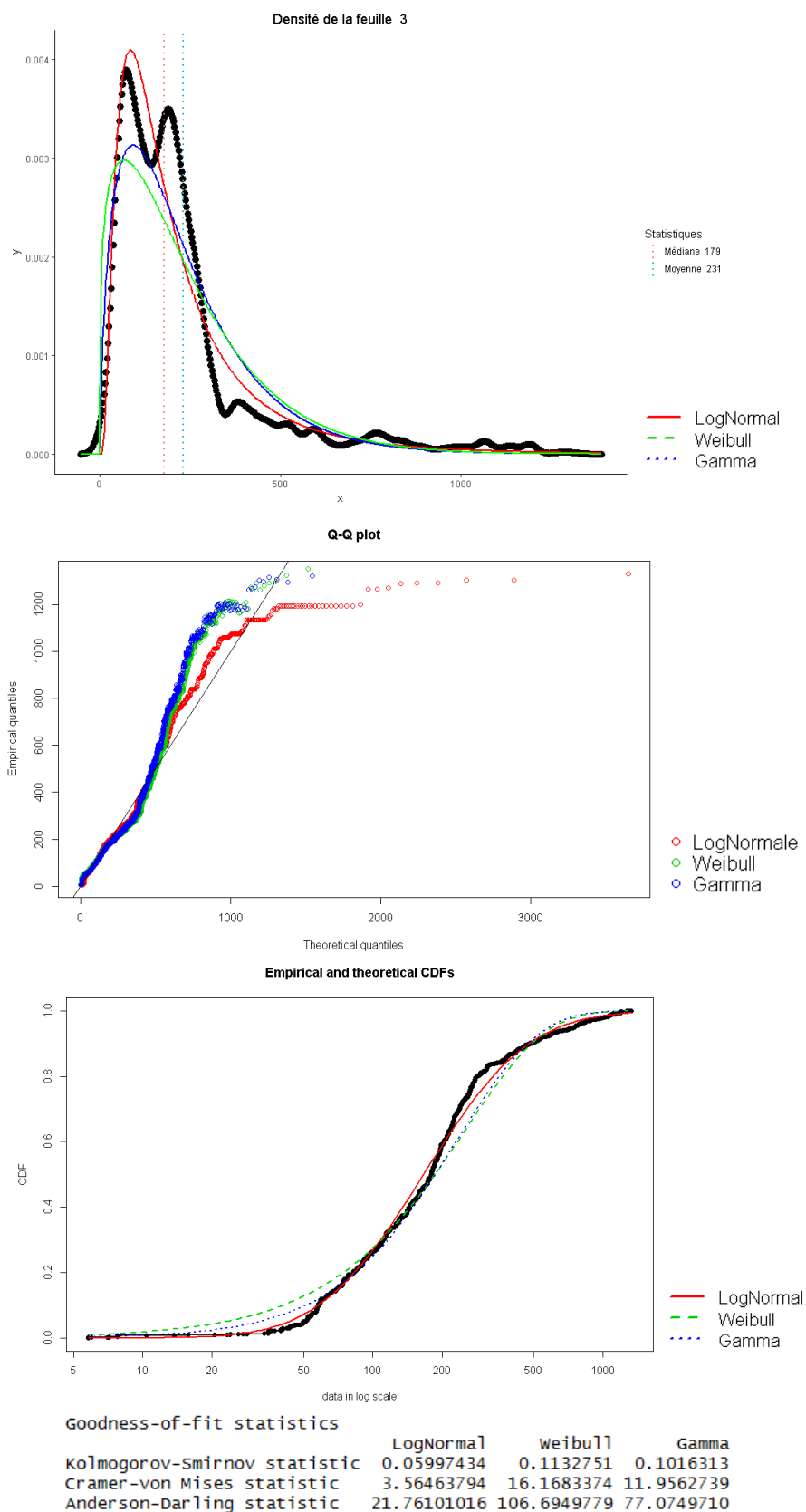


Figure B.29 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 3

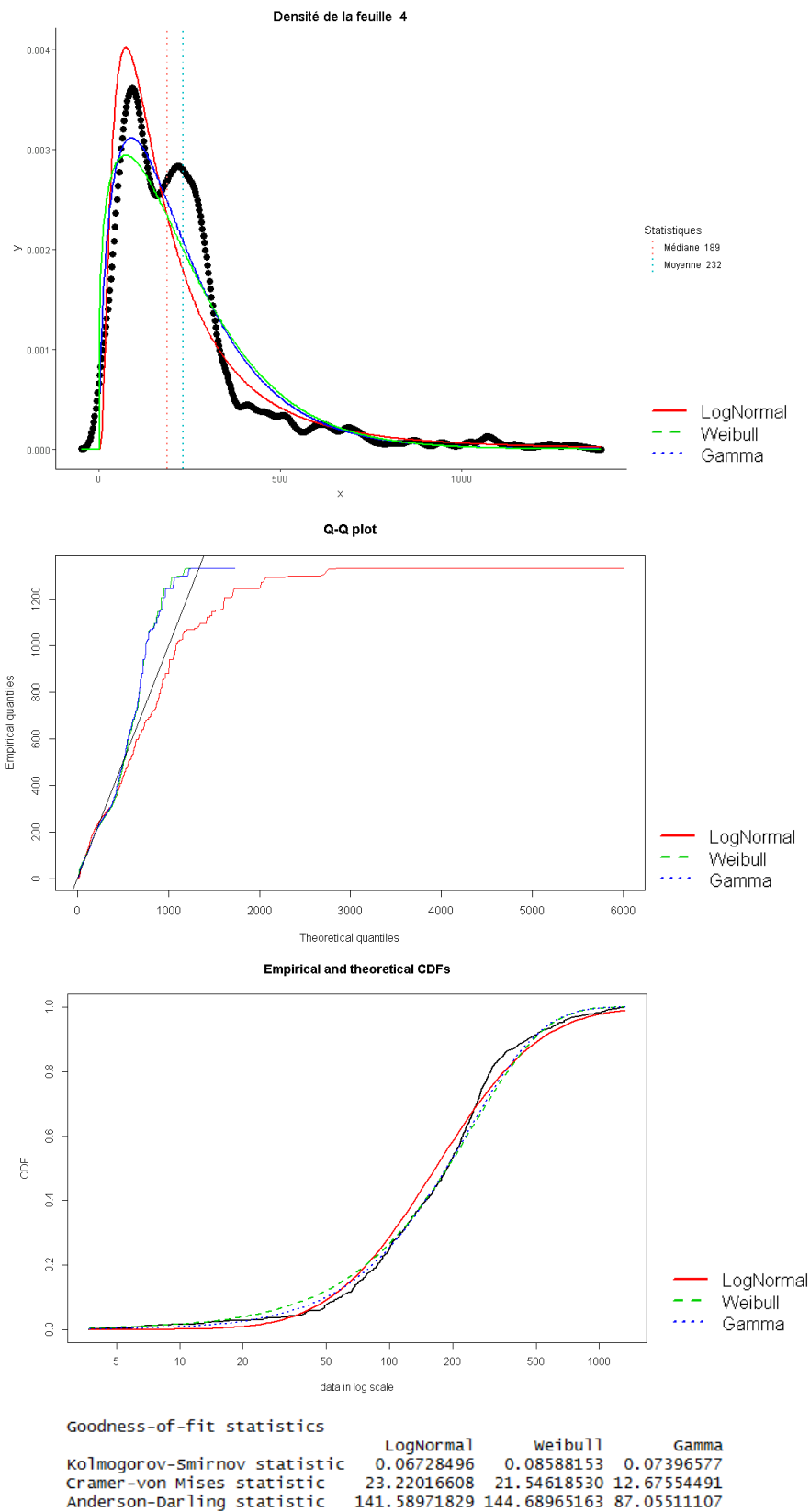


Figure B.30 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 4

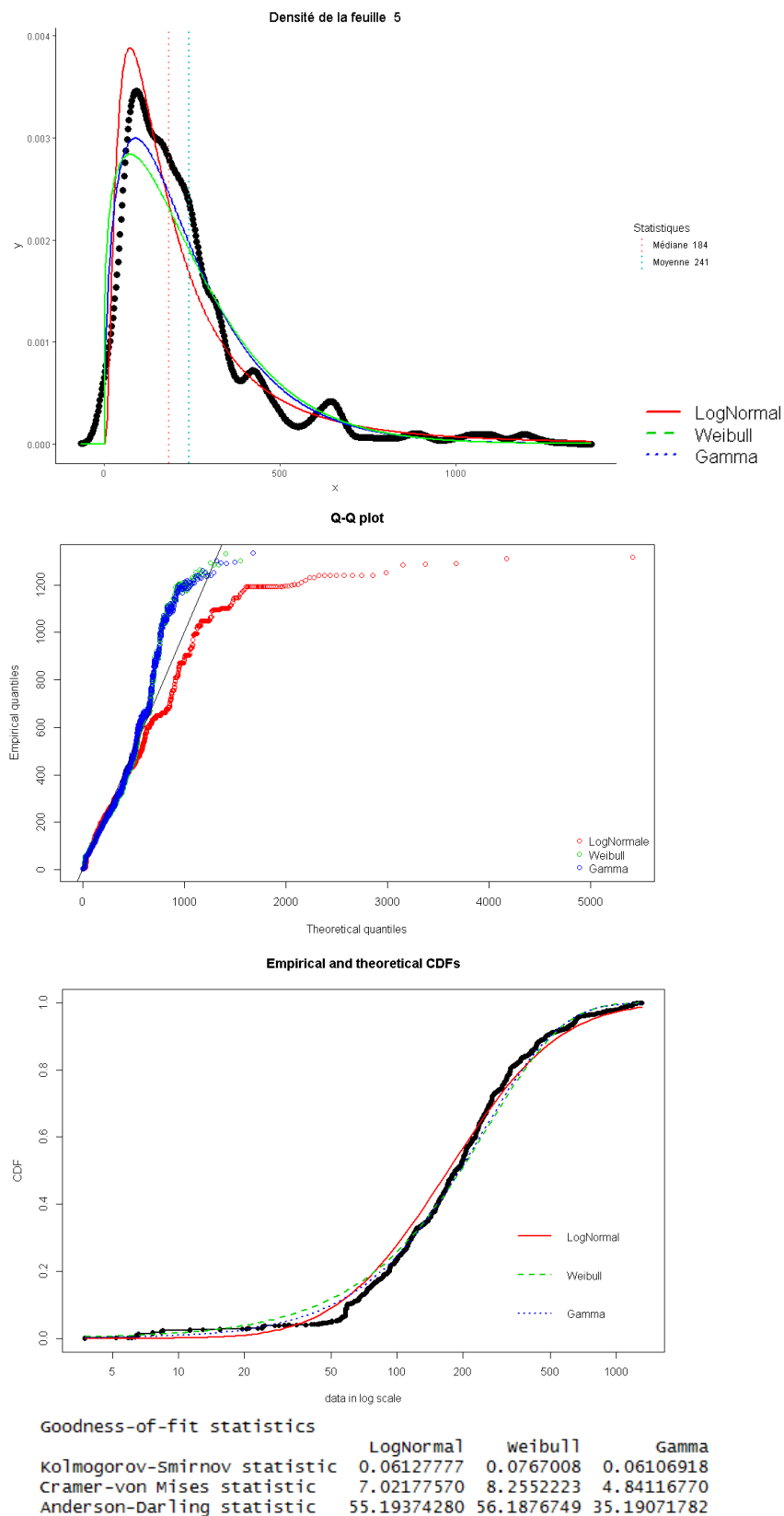


Figure B.31 – Graphiques et statistiques d’ajustement des lois de probabilités de la Feuille 5

Annexe C

Lien entre la prime et la valeur d'un contrat

Nous avons vu dans la **partie 1.1.1.2** la formule de la valeur d'un contrat. Il est possible d'écrire une telle formule autrement.

$$\text{En effet : } P^i = P^{i-1} \times K_i = P^1 \prod_{k=2}^i K_k$$

$$\text{Où : } K_i = TX_{MAJO}^i \times (1 - TX_{CHUTE}^i)$$

De la même façon, il est possible de réécrire les i-ème sinistre, frais, provision et marge de solvabilité en fonction de la prime initiale :

$$S^i = S^{i-1} \times L_i = S^1 \prod_{k=2}^i L_k = P^1 \times SP \times \prod_{k=2}^i L_k$$

$$\text{Où : } L_i = TX_{INFLA}^i \times (1 - TX_{CHUTE}^i)$$

$$F^i = P_i \times TX_{COMM} = P^1 \prod_{k=2}^i K_k \times TX_{COMM}$$

$$Provision^i = \sum_{k=1}^i TX_{PSAP}^k \times S^{i-k+1} = P_1 \times SP \sum_{k=1}^i TX_{PSAP}^k \times \prod_{k=2}^{i-k+1} L_k$$

$$MS^i = P^i \times TX_{MargeSolva} \times Coût_capital = P^1 \prod_{k=2}^i K_k \times TX_{MargeSolva} \times Coût_capital$$

Ainsi nous voyons que la valeur d'un contrat, étant ses caractéristiques, ne dépend que de la prime. Il est donc possible de calculer une *value-in-force* (VIF) de sorte que :

$$\text{Valeur du contrat} = P^1 \times VIF$$

Une fois le VIF calculé, il est possible de déterminer le rabais correspondant à une baisse x de valeur.

$$\text{Rabais sur la prime} = P_{HR} - \frac{\text{Valeur du contrat hors rabais} - x}{VIF}$$

Où : P_{HR} est la prime hors rabais.

Table des figures

1	À gauche : la régression logistique, à droite : le gradient boosting . . .	5
2	Probabilité de souscription des produits équipés	6
3	Left : logistic regression, right : gradient boosting	10
4	Probability of underwriting equipped products	11
1.1	Profits par année d'un contrat d'assurance vie Source : Mathilde Garotin, <i>Market-consistent embedded value</i> (MCEV), ISUP, 2018-2019	21
1.2	Décomposition du rendement d'un contrat par année de projection . .	23
1.3	Le modèle actuel de la valeur potentielle	26
1.4	Le modèle développé dans le mémoire	27
2.1	Répartition de la souscription d'un second contrat en fonction de l'an- cienneté	29
2.2	Evolution de l'équipement des clients par ancienneté	30
2.3	Répartition du produit d'appel parmi les clients professionnels	31
2.4	Taux de multi-équipement à la fin de la première année en portefeuille	33
2.5	Répartition du multi-équipement en fonction du produit d'appel . . .	34
2.6	Répartition des produits qui sont multi-équipés en fonction du produit d'entrée	35
2.7	Pourcentage de variables correctement renseignées	37
2.8	Variables découpées en classes	38
2.9	Variables découpées en classes	39
2.10	Taux moyen de clients qui se multi-équipent en fonction de la variable <i>Code Territoire</i>	40
2.11	Taux moyen de clients qui se multi-équipent en fonction de la variable <i>Type du réseau de distribution</i>	40
2.12	Taux moyen de clients qui se multi-équipent en fonction de la variable <i>Indicatrice du Statut de l'occupant</i>	41
2.13	Taux moyen de clients qui se multi-équipent en fonction de la variable <i>Indicatrice du Multi-équipement</i>	42
2.14	Taux moyen de clients qui se multi-équipent en fonction de la variable <i>Indicatrice de la possession d'au moins un contrat Automobile</i>	42
2.15	Taux moyen de clients qui se multi-équipent en fonction de la variable <i>Indicatrice de la CSP</i>	43
2.16	Taux moyen de clients qui se multi-équipent en fonction de la variable <i>Indicatrice de la banque du client</i>	43
2.17	Taux moyen de clients qui se multi-équipent en fonction de la variable <i>Indicatrice de la nullité de la prime IARD</i>	44

2.18	Taux moyen de clients qui se multi-équipent en fonction de la variable <i>Famille du produit d'appel (ou d'entrée)</i>	44
2.19	Taux moyen de clients qui se multi-équipent en fonction de la variable <i>Indicatrice de nombre d'enfants</i>	45
2.20	Taux moyen de clients qui se multi-équipent en fonction de la variable <i>Nombre de contrats</i>	45
2.21	Taux moyen de clients qui se multi-équipent en fonction de la variable <i>Composition du panier produit</i>	46
2.22	Taux moyen de clients qui se multi-équipent en fonction de la variable <i>Délai entre la souscription du premier et du second contrat</i>	47
2.23	Taux moyen de clients qui se multi-équipent en fonction de la variable <i>Délai entre la souscription du dernier et de l'avant-dernier contrat</i>	47
2.24	Diagrammes en boîte de la valeur des contrats en fonction des variables <i>Nature du client, Age et Réseau de distribution</i>	51
2.25	Diagrammes en boîte de la valeur des contrats en fonction des variables <i>Zoniers</i>	52
2.26	Diagrammes en boîte de la valeur des contrats en fonction des variables <i>Réseau de distribution et Nature de client</i>	52
2.27	Diagrammes en boîte de la valeur des contrats en fonction des variables <i>Nature de client, Age et Réseau de distribution</i>	53
2.28	Diagrammes en boîte de la valeur des contrats en fonction des variables <i>Zoniers</i>	53
3.1	Contrainte Lasso en bleu et fonction de Coût en rouge Source : Tibshirani Robert, Hastie Trevor et Wainwright Martin, <i>Statistical Learning with Sparsity : The Lasso and Generalizations</i> , Chapman and Hall, 2015	58
3.2	Contrainte Ridge en bleu et fonction de Coût en rouge Source : Tibshirani Robert, Hastie Trevor et Wainwright Martin, <i>Statistical Learning with Sparsity : The Lasso and Generalizations</i> , Chapman and Hall, 2015	59
3.3	Types de structures de données	60
3.4	Arbre de classification	61
3.5	Exemple : ensemble de données divisé en sous-espace et arbre de décision associé	62
3.6	Contribution des variables explicatives dans un modèle d'apprentissage Source : Lundberg Scott, « A unified approach to explain the output of any machine learning model » [en ligne], disponible sur https://github.com/slundberg/shap [consulté le 8 septembre 2019]	67
3.7	Schéma explicatif du la précision et du rappel (à gauche) et courbe ROC (à droite). Sources : Walber, 2014 (gauche) et Ch.-A. Azencott, Open-classrooms.com, 2019 (droite).	70
3.8	Performance du modèle en fonction de la pénalisation	73
3.9	Performances en fonction du nombre de variables sélectionnées <i>Courbe bleue : train. Courbe verte : test.</i>	73
3.10	Coefficients du modèle à 23 variables	74
3.11	Test du rapport de vraisemblance	74
3.12	Comparaison des prédictions et des valeurs réelles	75
3.13	Comparaison des prédictions et des valeurs réelles	77
3.14	Comparaison des prédictions et des valeurs réelles	77

3.15	Moyenne des performances sur l'échantillon de test en fonction de la profondeur des arbres et du nombre d'arbres lors de la validation croisée	79
3.16	Importance des variables pour les hyperparamètres choisis	79
3.17	Comparaison des prédictions et des valeurs réelles	81
3.18	Courbe ROC sur les données de 2017-2018	82
3.19	Comparaison des prédictions et des valeurs réelles	82
3.20	Aire sous la courbe ROC des scores de multi-équipement actuels	83
4.1	Répartition des produits souscrits par les clients multi-équipés et leur valorisation	87
4.2	Répartition des produits souscrits par les clients multi-équipés et valorisation - Suite	87
4.3	Probabilité de souscription des produits équipés en fonction de la classe du client	90
4.4	Répartition de la valeur hors offre commerciale des contrats <i>Automobile hors E.S.</i>	91
4.5	Lecture d'un arbre de décision	92
4.6	Arbre de régression pour la valeur des produits <i>Automobile hors E.S.</i>	92
4.7	Répartition de la valeur hors offre commerciale des contrats <i>Engins Spéciaux</i>	93
4.8	Arbre de régression pour la valeur des produits <i>Engins Spéciaux</i>	93
4.9	Répartition de la valeur hors offre commerciale des contrats <i>MRH & PNO</i>	94
4.10	Arbre de régression pour la valeur des produits <i>MRH & PNO</i>	95
4.11	Répartition de la valeur au sein de chaque feuille de l'arbre de régression pour les produits <i>Automobile hors E.S.</i>	97
4.12	Densités de la empirique et des lois ajustées	98
4.13	Diagramme quantile-quantile	99
4.14	Fonctions de répartition empirique et des lois ajustées	99
4.15	Statistiques d'ajustement	100
4.16	Tableau récapitulatif des lois de probabilité choisies pour chaque feuille	101
4.17	Densité de la valeur en cas de souscription	102
4.18	Quartiles et moyennes des distributions	103
5.1	Histogramme de la prime et de la valeur des contrats	106
5.2	Histogramme de la probabilité prédite et de la valeur moyenne en cas de souscription des clients	106
5.3	Statistiques descriptives liées à la valeur du portefeuille (en milliers) à la fin de la première année pour le <i>Scénario témoin</i>	107
5.4	Statistiques descriptives liées à la valeur du portefeuille (en milliers) à la fin de la première année pour le <i>Scénario 1</i>	108
5.5	Statistiques descriptives liées à la valeur du portefeuille (en milliers) à la fin de la première année pour le <i>Scénario 2</i>	109
5.6	Statistiques descriptives liées à la valeur du portefeuille (en milliers) à la fin de la première année pour le <i>Scénario 3</i>	110
5.7	Statistiques descriptives liées à la valeur du portefeuille (en milliers) à la fin de la première année pour le <i>Scénario 4</i>	111

A.1	Erreur de la prédiction de la probabilité pour un seuil de confiance à 95 %	117
B.1	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 1	119
B.2	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 2	120
B.3	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 3	121
B.4	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 4	122
B.5	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 5	123
B.6	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 6	124
B.7	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 7	125
B.8	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 8	126
B.9	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 9	127
B.10	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 10	128
B.11	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 11	129
B.12	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 12	130
B.13	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 1	131
B.14	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 2	132
B.15	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 3	133
B.16	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 4	134
B.17	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 5	135
B.18	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 6	136
B.19	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 7	137
B.20	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 8	138
B.21	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 9	139
B.22	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 10	140

B.23	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 11	141
B.24	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 12	142
B.25	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 13	143
B.26	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 14	144
B.27	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 1	145
B.28	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 2	146
B.29	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 3	147
B.30	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 4	148
B.31	Graphiques et statistiques d'ajustement des lois de probabilités de la Feuille 5	149