



المندوبية السامية للتخطيط  
HAUT-COMMISSARIAT AU PLAN

ROYAUME DU MAROC  
\*\_\*\_\*\_\*\_\*  
HAUT COMMISSARIAT AU PLAN  
\*\_\*\_\*\_\*\_\*\_\*\_\*\_\*\_\*\_\*

INSTITUT NATIONAL  
DE STATISTIQUE ET D'ECONOMIE APPLIQUEE



**INSEA**

**Projet de Fin d'Etudes**

\*\*\*\*\*

## **Elaboration d'un modèle de scoring appliqué à la microfinance**

Préparé par : *M. Rachad EL MROUJI*

Sous la direction de : *M. Abdelaziz CHAOUBI (INSEA)*  
*Mme Roukia LAHLOU (BCP CONSULTING)*  
*M. Hamza BENFDIL (BCP CONSULTING)*

*Soutenu publiquement comme exigence partielle en  
vue de l'obtention du*

**Diplôme d'Ingénieur d'Etat**

**Filière : ACTUARIAT-FINANCE**

*Devant le jury composé de :*

- *M. Abdelaziz CHAOUBI (INSEA)*
- *Mme Fadoua BADAoui (INSEA)*
- *Mme Roukia LAHLOU (BCP CONSULTING)*
- *M. Hamza BENFDIL (BCP CONSULTING)*

SEPTEMBRE 2020 / PFE N° 8



# Résumé

La microfinance joue un rôle socio-économique très important surtout dans les pays en cours de développement. Elle offre aux personnes pauvres ou à faible revenus la possibilité de bénéficier des différents services financiers tels que les crédits et l'épargne. A travers des microcrédits, ces individus auront accès à un capital qui leur permettrait de lancer des micro-activités génératrices de revenu.

Ces individus sont exclus par le système financier classique et notamment les banques de détail à cause de leurs incapacités de remboursement. En effet, ils sont classés comme risqués et incapables d'honorer leurs engagements à cause de leurs revenus faibles et instables, ainsi que l'absence des garanties matérielles. Autrement dit, il représente un risque de défaut très élevé.

Le risque de défaut lié à l'octroi des crédits représente un souci majeur pour les banques ainsi que pour les régulateurs. La réglementation baloise a défini plusieurs outils permettant la gestion de ce type de risques tels que le Credit Scoring.

Le Credit Scoring est un outil de notation des clients qui permet la prédiction de la performance des clients avant de leurs octroyer le crédit. L'objectif de ce mémoire est de développer un modèle de Scoring des prêts individuels pour une institution de microfinance en Côte d'Ivoire. Pour ce faire, on propose quatre modèles : la régression logistique, les arbres de classification, les forêts aléatoires et finalement le Gradient Boosting.

Les résultats de cette étude montrent que les modèles d'agrégation en machine learning -Random Forest et Gradient Boosting- sont les plus performants, ces modèles s'exécutent en boîte noire, d'où la difficulté de les interpréter ou de les intégrer dans le système d'information du client. Par conséquent, on choisit le modèle de la régression logistique, à partir duquel on construit une grille de notation selon le niveau du risque.

**Mots Clés : Microfinance, Scoring, Crédit, Probabilité de défaut, Régression logistique, Arbre de régression et de classification CART, Forêt aléatoire, Bootsrap, Gradient Boosting, Apprentissage Automatique**

# Abstract

Microfinance plays a very important socio-economic role, especially in developing countries. It offers poor and low-income people the opportunity to benefit from various financial services such as loans and savings. Through microcredits, these individuals will have access to capital that would allow them to launch income-generating micro-activities.

Those people are excluded by the traditional financial system especially retail banks because of their inability to repay loans. In fact, they are classified as risky and incapable to honor their commitments, which is because of their low and unstable income, and also, they do not have sufficient material guarantees. In other words, it represents a very high risk of default.

Credit risk linked to granting loans represents a major concern for banks as well as for regulators. The Basel regulations have defined several tools allowing the management of this type of risks such as Credit Scoring.

Credit Scoring is a tool that helps to predict a customer performance before granting him the loan. The objective of this thesis is to develop an individual loan scoring model for microfinance in Ivory Coast. For that, four models are proposed in this study: logistic regression, Classification And Regression Trees, Random forests and Gradient Boosting

The results of this study show that the machine learning aggregation models -Random Forest and Gradient Boosting- are the most efficient. These models are run in black box, which make it difficult to interpret or integrate into client's information system. Consequently, we choose the logistic regression model, from which we construct a scoring grid according to the level of risk.

**Mots Clés : Microfinance, Scoring, Credit, Probabilty of default, Logistic regression, Classification And Regression Tree CART, Random Forest, Bootsrap, Gradient Boosting, Machine Learning**

# Dédicace

A mes parents pour tous leurs sacrifices, leur amour, leur soutien et leurs prières tout au long de mes études.

A mes sœurs, mon frère et L pour leurs encouragements et leur soutien moral

A mes amis, le groupe CN, qui m'ont toujours aidé tout au long de mon parcours à l'INSEA

# Remerciements

Tout d'abord, Je tiens à remercier mes encadrants Mme Roukia LAHLOU et M. Hamza BENFDIL pour leurs confiances et surtout pour leurs accompagnements exceptionnels durant cette période difficile qu'a connu le Maroc et le monde entier.

Je tiens à remercier également le reste de l'équipe du pôle Modélisation et ingénierie financière : M. Reda JABRAZKO et M. Abdlouahab AGOUMI pour leurs conseils et disponibilités.

Je tiens à remercier Mon encadrant académique M. Abelaziz Chaoubi pour m'avoir encadré durant ce projet.

Et finalement, un grand merci à Mme Fadoua Badaoui pour avoir accepté d'évaluer mon travail.

## Table des matières

Résumé.....	3
Abstract.....	4
Dédicace.....	5
Remerciements .....	6
Table des matières.....	7
Liste des figures.....	10
Liste des tableaux .....	12
Liste des Abréviations.....	13
Partie I : Réglementation et notions.....	14
<b>1. Risque de crédit sous Bâle II.....</b>	<b>15</b>
<b>1.1. Risque de crédit .....</b>	<b>15</b>
<b>1.2. L’architecture de Bâle II .....</b>	<b>16</b>
<b>1.3. La défaillance.....</b>	<b>16</b>
<b>1.4. La notation des clients .....</b>	<b>17</b>
<b>1.5. Les composantes du risque de crédit.....</b>	<b>17</b>
<b>2. Credit Scoring .....</b>	<b>18</b>
<b>2.1. Définition .....</b>	<b>18</b>
<b>2.2. Les avantages du Credit Scoring .....</b>	<b>18</b>
<b>2.3. Les faiblesses du Credit Scoring.....</b>	<b>18</b>
<b>2.4. La démarche de construction d’un modèle de Scoring.....</b>	<b>19</b>
<b>3. Microfinance .....</b>	<b>20</b>
<b>3.1. Contexte .....</b>	<b>20</b>
<b>3.2. Différence entre microfinance et inclusion financière .....</b>	<b>20</b>
<b>3.3. Le microcrédit .....</b>	<b>21</b>
<b>3.4. Microfinance en Côte d’Ivoire .....</b>	<b>21</b>
<b>4. Objectif et étapes de l’étude.....</b>	<b>24</b>
Partie II : Cadre théorique .....	25
<b>1. La régression logistique .....</b>	<b>26</b>
<b>1.1. Les limites de la régression linéaire.....</b>	<b>26</b>
<b>1.2. Odds.....</b>	<b>27</b>
<b>1.3. Fonction de vraisemblance .....</b>	<b>30</b>
<b>1.4. Propriétés des estimateurs.....</b>	<b>30</b>
<b>1.5. Méthode de sélection des variables .....</b>	<b>31</b>
<b>2. Arbre de classification.....</b>	<b>32</b>

2.1.	Définition .....	32
2.2.	Construction d'un arbre de classification .....	33
2.3.	Exemple .....	34
2.4.	Problème de surajustement .....	36
3.	Forêt aléatoire .....	39
3.1.	Bootstrap.....	39
3.2.	Construction d'une forêt aléatoire .....	39
3.3.	Out Of Bag.....	41
3.4.	Prédiction.....	41
4.	Gradient boosting.....	42
4.1.	Définition .....	42
4.2.	Arbre de régression CART .....	42
4.3.	Fonction de perte .....	44
4.4.	Algorithme .....	45
5.	Comparaisons de la qualité des modèles .....	49
5.1.	Séparation de bases : Apprentissage / test .....	49
5.2.	Cross Validation .....	49
5.3.	Matrice de confusion .....	50
5.4.	Courbe de ROC et AUC .....	50
Partie III : Application & Modélisation .....		51
1.	Traitement de la base de données .....	52
2.	Analyse exploratoire des variables .....	56
2.1.	Variables quantitatives .....	56
2.2.	Variables qualitatives .....	58
2.2.1.	Corrélations .....	58
2.2.2.	Effets sur le Défaut.....	58
3.	Modélisation .....	62
3.1.	Régression logistique .....	62
3.1.1.	Premier modèle .....	62
3.1.2.	Modèle optimal .....	63
3.1.3.	Les effets des variables sur la probabilité de défaut .....	64
3.1.4.	Détermination du seuil de prédiction .....	66
3.2.	Arbre de classification.....	68
3.2.1.	Construction de l'arbre .....	68
3.2.2.	Estimation des probabilités de défaut .....	68
3.2.3.	Analyse de la performance sur la base d'apprentissage .....	69

3.3.	Forêt aléatoire.....	70
3.3.1.	Choix du nombre d'arbre dans le modèle.....	70
3.3.2.	Choix du nombre de variables à considérer dans chaque répartition.....	70
3.3.3.	Effets des variables .....	71
3.3.4.	Probabilités de défaut estimées.....	72
3.3.5.	Performance du modèle sur la base d'apprentissage.....	72
3.4.	Gradient Boosting.....	73
3.4.1.	Choix des paramètres du modèle.....	73
3.4.2.	Les effets des variables sur le modèle .....	74
3.4.3.	Probabilité de défaut.....	75
3.4.4.	Performance du modèle sur la base d'apprentissage.....	75
4.	Comparaisons des modèles .....	78
4.1.	Base d'apprentissage .....	78
4.2.	Validation sur la base test .....	78
4.3.	Cross Validation .....	79
4.3.1.	Régression logistique.....	79
4.3.2.	Arbre de décision .....	82
4.3.3.	Random Forest.....	83
4.3.4.	Gradient Boosting .....	85
4.3.5.	Comparaison des modèles.....	88
4.4.	Choix du modèle .....	89
4.4.1.	Choix du modèle .....	89
4.4.2.	Grille de notation .....	90
	Conclusion.....	92
	Bibliographie.....	93
	Annexe .....	94
1.	Annexe 1 : Bibliothèques R.....	94
2.	Annexe 2 : Cross Validation .....	94
2.1.	Arbre de Classification.....	94
2.2.	Random Forest .....	99
2.3.	Gradient Boosting.....	100

# Liste des figures

Figure 1 : Processus de Credit Scoring .....	18
Figure 2 : Nombre de clients d'un échantillon de 20 IMF en Côte d'Ivoire .....	22
Figure 3 : Encours des crédits (en millions FCFA) d'un échantillon de 20 IMF en Côte d'Ivoire.....	22
Figure 4 : Montant des dépôts (en millions FCFA) d'un échantillon de 20 IMF en Côte d'Ivoire .....	23
Figure 5 : comparaison des régressions linéaires de la variable défaut en modifiant la codification ....	26
Figure 6 : Modélisation de la probabilité de défaut par un sigmoïde.....	27
Figure 7 : Asymétrie des Odds .....	28
Figure 8 : Distribution de la fonction logit : simulation sur 1000 probabilités .....	28
Figure 9 : Transformation de la variable défaut par la fonction logit .....	29
Figure 10 : Principe d'utilisation de la vraisemblance .....	29
Figure 11 : Composantes d'un arbre de classification .....	32
Figure 12 : Variation de l'indice de diversité de Gini .....	34
Figure 13 : Exemple pour la construction d'un arbre de classification .....	34
Figure 14 : Calcul indice de Gini pour une variable qualitative .....	36
Figure 15 : Problème de surajustement.....	37
Figure 16 : Principe du Bootstrap.....	39
Figure 17 : Algorithme de construction d'une forêt aléatoire .....	40
Figure 18 : Schéma du processus du Random Forest .....	40
Figure 19 : Arbre de régression : Choix du seuil de séparation pour une variable quantitative .....	43
Figure 20 : Arbre de régression : Choix de la meilleure combinaison de séparation pour une variable qualitative.....	43
Figure 21 : Fonctionnement d'une fonction de perte.....	44
Figure 22 : Algorithme de construction d'un modèle GBM.....	45
Figure 23 : Illustration de la technique de Cross Validation .....	49
Figure 24 : Distribution des variables qualitatives .....	56
Figure 25 : matrice des corrélations des variables quantitatives.....	58
Figure 26 : Boîtes à moustache des variables quantitatives en fonction du défaut.....	58
Figure 27 : Premier modèle de RL.....	62
Figure 28 : Modèle optimale de la RL .....	63
Figure 29 : Courbe de ROC de la RL sur la base d'apprentissage .....	63
Figure 30 : Courbes de ROC des RL des variables du modèle su le défaut .....	64
Figure 31 : Courbes des effets des variables sur le défaut .....	65
Figure 32 : probabilités de défaut estimées par la RL.....	66
Figure 33 : les erreurs de performance de la RL sur la base d'apprentissage .....	67
Figure 34 : Arbre de décision obtenue par le modèle Arbre de classification .....	68
Figure 35 : Probabilités de défaut estimées pour l'arbre de classification.....	69
Figure 36 : Courbe de ROC de l'arbre de classification sur la base d'apprentissage .....	69
Figure 37 : les taux d'erreurs commises par le RF en fonction du nombre d'arbres .....	70
Figure 38 : Variation de l'erreur d'OBB en fonction du nombre de variables considérées dans chaque répartition .....	71
Figure 39 : Effets des variables sur le RF .....	71
Figure 40 : Probabilités de défauts estimées sur la base d'apprentissage pour le modèle RF .....	72
Figure 41: Courbe de ROC du modèle RF sur la base d'apprentissage.....	72
Figure 42 : sélection du nombre d'arbre optimal pour un taux d'apprentissage de 0.01 par la méthode de Cross Validation.....	73
Figure 43 : sélection du nombre d'arbre optimal pour un taux d'apprentissage de 0.1 par la méthode de Cross Validation .....	74
Figure 44 : Relative Influence des variables sur le GBM .....	74
Figure 45 : Probabilités de défaut prédites pour le modèle GBM sur la base d'apprentissage .....	75
Figure 46 : Courbe de ROC pour le modèle GBM sur la base d'apprentissage.....	76

Figure 47 : Performance du modèle GBM sur la base d'apprentissage.....	76
Figure 48 : Courbes de ROC des 4 modèles sur la base de test.....	78
Figure 49 : Courbes de ROC de la RL sur les bases test - Cross Validation 10 classes .....	80
Figure 50 : Courbes de ROC de l'arbre de classification CART sur les bases test - Cross Validation 10 classes .....	82
Figure 51 : Importance des variables pour le RF – Cross Validation .....	84
Figure 52 : Courbes de ROC de RF sur les bases test - Cross Validation 10 classes .....	84
Figure 53 : Influence des variables sur le GBM - Cross Validation .....	86
Figure 54 : Courbes de ROC de GBM sur les bases test - Cross Validation 10 classes .....	86
Figure 55 : Comparaison des courbes de ROC entre les 4 modèles - Cross Validation .....	88
Figure 56 : Taux de performance du RF en fonction des probabilités – Cross Validation .....	89
Figure 57 :: Taux de performance de la RL en fonction des probabilités – Cross Validation.....	90
Figure 58 : Arbre de Classification 1 - Cross Validation .....	94
Figure 59 : Arbre de Classification 2 - Cross Validation .....	95
Figure 60 : Arbre de Classification 3 - Cross Validation .....	95
Figure 61 : Arbre de Classification 4 - Cross Validation .....	96
Figure 62 : Arbre de Classification 5 - Cross Validation .....	96
Figure 63 : Arbre de Classification 6 - Cross Validation .....	97
Figure 64 : Arbre de Classification 7 - Cross Validation .....	97
Figure 65 : Arbre de Classification 8 - Cross Validation .....	98
Figure 66 : Arbre de Classification 9 - Cross Validation .....	98
Figure 67 : Arbre de Classification 10 - Cross Validation .....	99

# Liste des tableaux

Tableau 1 : Architecture de Bâle II.....	16
Tableau 2 : Matrice de confusion .....	50
Tableau 3 : variables utiles pour l'analyse, mais mal renseignées sur la base de données .....	53
Tableau 4 : Variables tirées de la base sans traitements.....	53
Tableau 5 : Variables traitées .....	55
Tableau 6 : résumé des tables de contingence des variables qualitatives.....	57
Tableau 7 ; tests de Student d'égalité des moyennes des variables qualitatives des deux populations : défauts et sains.....	61
Tableau 8 : matrice de confusion de la RL sur la base d'apprentissage pour une probabilité de 0,5 .....	67
Tableau 9 : Matrice de Confusion de l'arbre de classification sur la base d'apprentissage .....	70
Tableau 10 : Matrice de confusion du modèle RF sur la base d'apprentissage .....	73
Tableau 11 : Matrice de confusion du modèle GBM sur la base d'apprentissage pour une probabilité de 0.48 .....	76
Tableau 12 : Indicateurs de performance des 4 modèles sur la base d'apprentissage .....	78
Tableau 13 : Indicateurs de performance des 4 modèles sur la base de test.....	79
Tableau 14 : Estimations des paramètres de la RL sur les 10 classes de la Cross Validation .....	79
Tableau 15 : Indicateurs de la performance de la RL - Cross Validation 10 Classes .....	81
Tableau 16 : Indicateurs de la performance de l'arbre de classification CART - Cross Validation 10 Classes.....	83
Tableau 17 : Indicateurs de la performance du RF - Cross Validation 10 Classes.....	85
Tableau 18 : Indicateurs de la performance du GBM - Cross Validation 10 Classes.....	87

# Liste des Abréviations

IMF : Institution de microfinance

GBM : Gradient Boosting Machine

RF : Random Forest

RL : Régression logistique

CART: Classification And Regression Trees

PD : Probabilité de défaut

PCA : Perte en cas de défaut

ECD : Exposition en cas de défaut

ROC: Receiver operating characteristic

AUC: Area under the curve

VP : Vrais Positifs

VN : Vrais Négatifs

FP : Faux Positifs

FN : Faux Négatif

# **Partie I :**

# **Réglementation**

# **et notions**

## 1. Risque de crédit sous Bâle II

L'octroi du crédit représente la fonction principale de la banque, c'est une opération par laquelle cette dernière met à la disposition d'un client un montant d'argent pour une durée déterminée, en contrepartie, le client s'engage à verser des intérêts. Même si la banque espère dégager un profit de cette opération, elle s'expose simultanément à une incertitude de non remboursement de l'emprunteur.

Pour assurer une bonne performance, la banque est amenée à gérer ce risque de crédits. Dans ce contexte, La réglementation baloise propose des ratios relatifs à ce risque ainsi que des méthodes de calcul pour les estimer.

La réforme Bâle II comporte des modifications sur le ratio de solvabilité en incluant le risque opérationnel, et en introduisant de nouvelles approches de calcul des exigences concernant le risque de crédit.

### 1.1. Risque de crédit

Un risque se définit comme la possibilité de survenance d'un événement ayant des conséquences négatives. Dans le cas du risque de crédit, il correspond au risque d'une perte potentielle relative au défaut d'un emprunteur par rapport au remboursement de ses dettes (obligations, prêts bancaires, créances commerciales ...). En d'autres termes, il correspond à une situation, par bonne ou mauvaise foi, où l'emprunteur se trouve dans l'incapacité d'honorer ses engagements.

Par exemple, dans le cas des crédits classiques tels que les prêts bancaires ou les prêts obligataires, le risque de crédit se révèle en cas de survenance d'un problème pour un paiement prévu tel que : le non remboursement du capital et/ou des intérêts à la date prévue dans le contrat, le remboursement partiel à la date d'échéance, ou le report d'un paiement. Il peut provenir également du défaut d'une contrepartie.

## 1.2.L'architecture de Bâle II

Pilier I	Pilier II	Pilier II
Exigences minimales de fonds propres	Surveillance par les autorités de supervision	Transparence et discipline de marché
<ul style="list-style-type: none"> <li>• Risque de crédit (nouvelles approches) :               <ul style="list-style-type: none"> <li>○ Approche standard, basée sur les notations externes.</li> <li>○ Approche fondation de notation interne simple.</li> <li>○ Approche avancée de notation interne complexe.</li> </ul> </li> <li>• Risque opérationnel (nouveau).</li> <li>• Risque de marché (inchangé).</li> </ul>	<ul style="list-style-type: none"> <li>• L'analyse par la banque de ses risques non couverts par le pilier I (risque de liquidité, de taux, de concentration, stress tests ...) et la revue des actions qu'elle doit entreprendre pour gérer ses autres risques (fonds propres supplémentaires, provisions, actions de contrôle interne ou gestion des risques)</li> <li>• L'évaluation des mécanismes interne d'appréciation du niveau de fonds propres par les autorités de contrôle.</li> <li>• L'intervention des autorités de contrôle pour garantir que les banques respectent les règles et disposent des niveaux minimaux des fonds propres.</li> </ul>	<ul style="list-style-type: none"> <li>• Obligations accrues de publication (sur les fonds propres et les différentes méthodes d'évaluation des risques).</li> </ul>
<p>Bâle I :</p> <ul style="list-style-type: none"> <li>• Harmonisation des règles en matière d'exigence de capital.</li> <li>• Taux de capital de 8% fixé de manière uniforme à l'échelle internationale.</li> </ul>		

Tableau 1<sup>1</sup> : Architecture de Bâle II

## 1.3.La défaillance

Le défaut de la part d'un débiteur, selon l'accord de Bâle II d'Avril 2003, se définit lorsque l'un des deux événements suivants se produit :

<sup>1</sup> Source : Comité de Bâle sur le contrôle bancaire, Vue d'ensemble du Nouvel accord de Bâle sur les fonds propres, Avril 2003

- La banque estime improbable que le débiteur rembourse en totalité son crédit au groupe bancaire sans qu'elle ait besoin de prendre des mesures appropriées telles que la réalisation d'une garantie (si elle existe).
- L'arriéré du débiteur sur son crédit important dû au groupe bancaire dépasse 90 jours

#### 1.4. La notation des clients

La réglementation bâloise élabore une analyse basée sur la notation des prêts dans le but de les différencier selon leurs niveaux de risque ; la notation est une évaluation du risque de non remboursement en temps et en heure de la totalité du principal et des intérêts relatifs à une obligation financière ou du capital restant dû dans le cas d'un crédit bancaire classique ; autrement dit, il s'agit d'une évaluation de la probabilité de défaut à un horizon. Pour cela, l'accord Bâle II élabore essentiellement deux approches :

- Approche standardisée : cette approche est plus simple, la banque affecte des pondérations aux actifs en fonction des risques, ces pondérations sont basées sur des évaluations externes du crédit.
- Approche fondée sur les notations internes (NI) : La banque peut s'appuyer sur ses estimations internes des composantes du risque pour déterminer l'exigence de fonds propres en regard d'une exposition donnée.

#### 1.5. Les composantes du risque de crédit

Les approches de notation déjà citées, ont pour but d'estimer les paramètres qui permettent de calculer l'exigence de fonds propres pour ce risque :

- Probabilité de défaut (PD) : ce paramètre correspond à la probabilité qu'un actif ou un emprunteur tombe en défaut sur une période donnée.
- Exposition en cas de défaut (ECD) : cet indicateur représente le montant maximal la banque peut perdre en cas de défaut, dans le cas d'une obligation, il représente le nominal, et dans le cas d'un prêt bancaire classique, il représente l'encours.
- Perte en cas de défaut (PCD) : ce paramètre concerne le pourcentage de l'ECD qui est perdu en cas de défaut. En effet ; en cas de défaut, la banque peut recouvrir une partie de l'encours restant en utilisant la garantie attachée (si elle existe) par exemple.

Finalement, la relation concernant la perte de crédit attendue (PCA) est donnée par Bâle II par la relation suivante :  **$PCA = PD \times PCD \times ECD$** .

## 2. Credit Scoring

### 2.1. Définition

L'octroi des crédits est une opération risquée pour les banques, mais en même temps, elle représente l'une des plus grandes sources de profit. Idéalement, ces les banques préféreraient de n'autoriser des crédits qu'aux client qui représenteraient une capacité élevée de rembourser les sommes prêtées. Pour cela, elles essaient de pronostiquer la performance de ces clients à l'aide des différents outils tel que le Credit Scoring.

Le Credit Scoring est un outil de gestion des risques qui a comme but la prévision des probabilités de défaillance d'un crédit avant son octroi.

L'objectif du Credit Scoring est la prédiction de la performance des prêts futurs en se basant sur des mesures quantitatives de la performance des prêts précédents qui avaient les mêmes caractéristiques.

Le processus de cet outil peut être résumé par le schéma suivant :

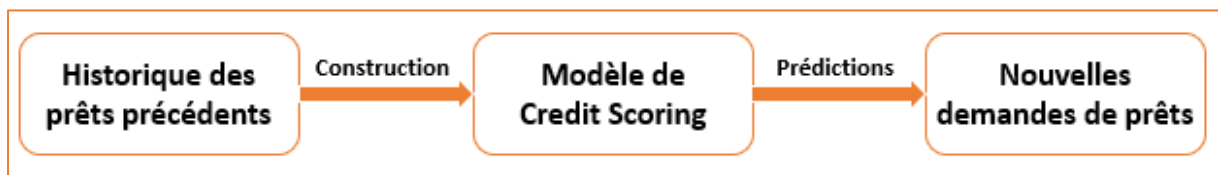


Figure 1 : Processus de Credit Scoring

### 2.2. Les avantages du Credit Scoring

- Il est objectif : au contraire de l'évaluation classique des crédits qui dépend des agents (expérience, humeur, ou discrimination par race ou religions ...), le modèle du Credit Scoring attribue des notations identiques aux clients qui représentent les mêmes caractéristiques.
- Il quantifie le risque sous forme d'une probabilité offrant une meilleure prédictibilité de la performance des prêts.
- Il améliore l'efficacité du processus d'analyse du crédit en optimisant le temps de l'analyse et les ressources nécessaires.
- Il tient en compte un grand nombre de facteurs de risques simultanément.
- Il est un outil d'aide à la décision pour la stratégie commerciale de la banque

### 2.3. Les faiblesses du Credit Scoring

- Il suppose que le futur réagira de la même façon que le passé
- Il nécessite un historique de bonne qualité sur un grand nombre de prêts

- Il nécessite également plusieurs données pour chaque prêt pour couvrir tous les aspects du risque
- Il est construit généralement sur les prêts acceptés par les agents, les crédits rejetés ne figurent pas dans les bases de données.
- Il suppose qu'une partie importante du risque est liée aux caractéristiques quantifiées
- Il peut réduire l'accès au crédit pour ceux qui n'ont pas d'historique

#### **2.4.La démarche de construction d'un modèle de Scoring**

La construction d'un modèle de Scoring passe par plusieurs étapes : Premièrement, il faut définir le défaut (retard de paiement de plus de 90 jours, des problèmes judiciaires ou une faillite ...), puis construire un échantillon qui contient les deux groupes d'emprunteurs : les "bons" profil qui n'ont pas enregistré le défaut ainsi que les "mauvais". Deuxièmement, analyser les déterminant du défaut pour sélectionner les variables discriminantes, ces variables doivent toucher les différents aspects du risque. Finalement, construire le modèle à l'aide d'une technique statistique (par exemple la régression logistique), puis élaborer une règle de décision, c'est-à-dire qui permettent la classification des clients, ensuite tester les résultats obtenus pour évaluer la performance précatrice du modèle.

## 3. Microfinance

### 3.1. Contexte

Historiquement, la microfinance référait surtout au microcrédit. Ce dernier correspond à l'attribution des crédits de faible montant à des personnes ayant un peu de revenu, à des entrepreneurs ou à des artisans qui ne peuvent pas accéder au système bancaire classique.

Ces banques n'accordent pas de tels crédits principalement pour trois raisons :

- Bien évidemment les coûts de gestion de plusieurs crédits sont plus élevés qu'un seul indépendamment des montants. En outre, le total des profits dégagés par un ensemble de crédits de faibles montants peut être équivalent au profit dégagé d'un crédit de grand montant.
- La deuxième raison concerne la pauvreté des clients. Ces clients ne possèdent pas assez de garanties matérielles.
- Asymétrie de l'information entre la banque et l'emprunteur qui résulte de l'incapacité de la banque à évaluer correctement la demande de financement à cause d'absence de données ou bien son incapacité d'observer les performances de cet emprunteur<sup>2</sup>.

Pour ces raisons, les banques se trouvent incapables de travailler correctement avec ce type de clients et préfèrent de les ignorer. Par conséquent, il fallait inventer des nouvelles politiques afin de les inclure dans le système financier.

### 3.2. Différence entre microfinance et inclusion financière

La différence principale entre les deux concepts se trouve dans leurs objectifs implicites<sup>3</sup>. L'objectif de l'inclusion financière est purement quantitatif qui vise à avoir le 100% d'inclusion bancaire. En d'autres termes, elle vise à ce que tous les individus utilisent un compte bancaire. Par contre, la microfinance ne vise pas seulement l'accès au système financier mais également l'impact, autrement dit, la capacité d'améliorer la vie des individus par la finance.

La microfinance était l'acteur principal dans l'inclusion financière, les institutions de microfinance (IMF) étaient les seules à opérer dans les régions isolées. Mais ce rôle commence à diminuer avec l'apparition de nouveaux services tels que les fintechs et le Mobile Banking. Ces acteurs permettent aux individus, et notamment ceux qui habitent dans les régions isolées, l'accès au système bancaire avec des coûts faibles, ils ont les moyens de toucher à ces clients sans construire du bâtiment dans ces régions.

---

<sup>2</sup> M.Goyer (1995)

<sup>3</sup> Renée Chao-Beroff

### 3.3. Le microcrédit

Le microcrédit consiste à prêter des faibles montants à des personnes n'ayant pas accès au système bancaire classique. Le microcrédit sert essentiellement à développer des activités génératrices de revenus.

Les caractéristiques principales du microcrédit :

- Montant : le montant d'un microcrédit est plus faible que celui d'un prêt bancaire classique.
- Durée de remboursement : les durées sont également plus courtes que celles d'un crédit classique.
- Taux d'intérêt : les taux sont généralement plus élevés que ceux d'un prêt bancaire. C'est par ce que les IMF ont des frais supplémentaires qu'ils doivent couvrir tels que la gestion de plusieurs crédits à faibles montants et la construction des agences dans des régions isolées ...
- Garanties : contrairement aux banques, la plupart des IMF ne demandent pas de garantie matérielle à leurs clients.

Pour résoudre le problème des garanties, les IMF généralement comptent sur la solidarité d'un groupe d'emprunteurs. Ce type de crédit était développé par Muhammad Yunus dans les années 70. La logique derrière le crédit solidaire consiste à prêter la somme totale à un groupe de personnes, chacune investit dans son activité le montant dont elle avait besoin pour le crédit. La responsabilité de remboursement est partagée par tout le groupe, ce qui met une pression sociale sur les individus, car une défaillance d'une seule personne signifie la défaillance du groupe.

Le microcrédit n'est pas le seul service offert par la microfinance, elle fournit également un ensemble de produits financiers aux personnes exclues tels que l'épargne et les services d'assurance et de transfert d'argent.

### 3.4. Microfinance en Côte d'Ivoire

La microfinance en Côte d'Ivoire est en pleine expansion, elle attire de plus en plus des clients. Le taux d'utilisation des services de microfinance sur la base de la population des adultes passe de 6% en 2006 à 11% en 2018<sup>4</sup>.

Les publications de la BCEAO concernant un échantillon de 20 SFD<sup>5</sup> entre 2017 et 2020 montrent que le nombre de clients de cet échantillon augmente de 1 182 340 jusqu'à 2 270 926.

---

<sup>4</sup> Source : BCEAO

<sup>5</sup> SFD : Système Financier Décentralisé, appellation francophone des IMF

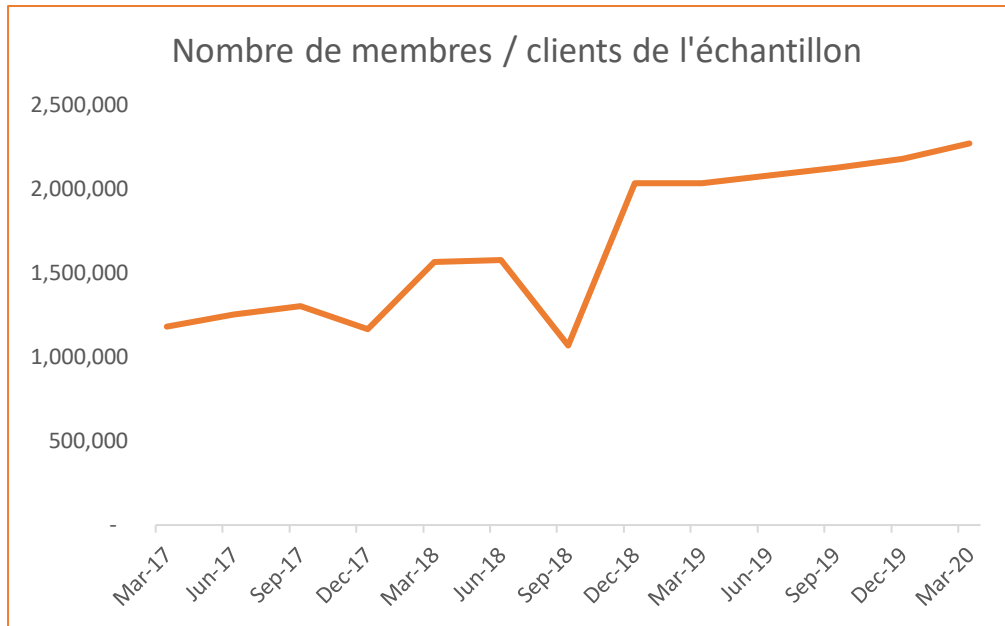


Figure 2 : Nombre de clients d'un échantillon de 20 IMF en Côte d'Ivoire<sup>6</sup>

L'encours des crédits de ces clients également augmente, il varie de 190 621 millions FCFA au début de 2017 à 323 599 millions FCFA au début de 2020.

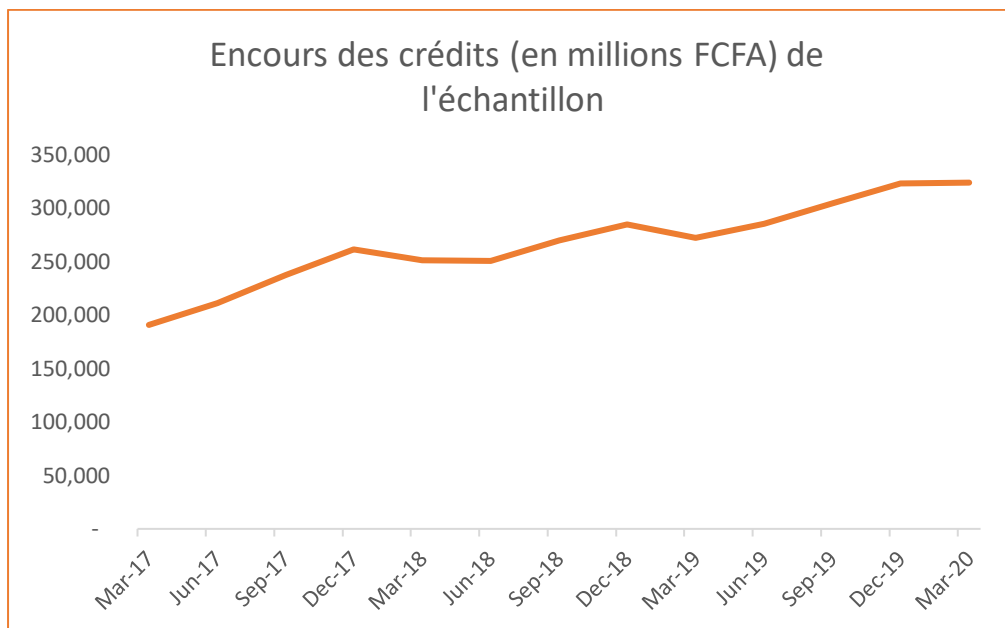


Figure 3 : Encours des crédits (en millions FCFA) d'un échantillon de 20 IMF en Côte d'Ivoire<sup>7</sup>

En ce qui concerne l'épargne, les montants déposés par la clientèle sont augmentés de 217 650 millions FCFA du début de 2017 à 315 129 millions FCFA au début de 2020.

<sup>6</sup> Source : BCEAO

<sup>7</sup> Source : BCEAO

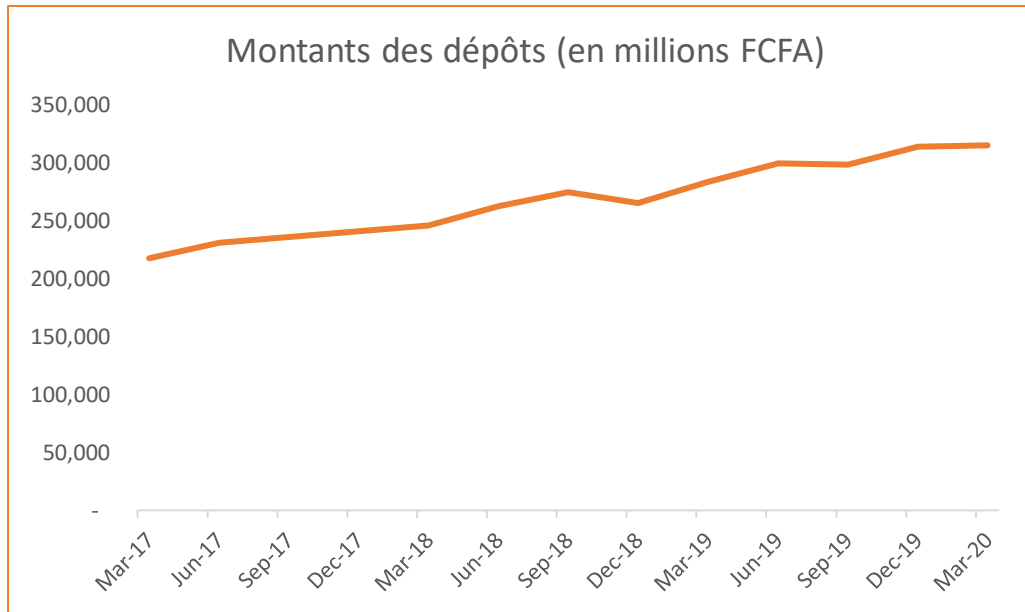


Figure 4 : Montant des dépôts (en millions FCFA) d'un échantillon de 20 IMF en Côte d'Ivoire<sup>8</sup>

Ces trois graphes nous montrent que les individus en Côte d'Ivoire ont de plus en plus tendance à utiliser les services de la microfinance.

<sup>8</sup> Source : BCEAO

## **4. Objectif et étapes de l'étude**

L'objectif principale de cette étude consiste à développer un outil de Scoring à l'octroi des crédits individuels pour une institution de microfinance en Côte d'Ivoire.

Pour atteindre cet objectif, nous suivrons les étapes suivantes :

- Traitement et analyse de la base de données
- Détermination du défaut adapté au contexte de l'IMF
- Choix des variables utiles pour la modélisation
- Modélisation à l'aide de plusieurs techniques
  - Modèle statistique : Régression logistique
  - Modèle d'apprentissage automatique : arbre de classification CART
  - Modèles d'apprentissage automatique par agrégation : Forêt aléatoire et Gradient Boosting
- Validation et choix du modèle
- Construction d'une grille de notation des clients.

# Partie II : Cadre théorique

# 1. La régression logistique

## 1.1. Les limites de la régression linéaire

Dans le cas d'une variable dépendante dichotomique, c'est-à-dire qui prend deux valeurs : Oui ou Non par exemple, ou qualitative en général, on ne peut pas appliquer la régression linéaire pour expliquer cette variable puisqu'elle ne vérifie pas les hypothèses de base de ce modèle.

Dans cette étude, la variable à expliquer est une variable binaire indiquant le défaut de remboursement, elle peut être recodifiée comme suit :

$$Y = \begin{cases} \text{Oui en cas de Défaut} \\ \text{Non sinon} \end{cases} \quad \text{ou} \quad Y = \begin{cases} 1 & \text{en cas de Défaut} \\ 0 & \text{sinon} \end{cases}$$

On pose  $X$  le vecteur des variables explicatives

Par construction, la variable  $Y$  suit une loi de Bernoulli :  $Y/X \sim \mathcal{B}(p(X))$ , avec une espérance :  $E[Y/X] = p(X)$

Un modèle de régression linéaire simple s'écrit de la façon suivante :  $E[Y/X] = X'\beta$  ou  $Y = X'\beta + e$ ; avec  $e$  vecteur des erreurs

$$\text{Par conséquent, } e = Y - X'\beta \text{ c\`ad } e_i = \begin{cases} 1 - x'_i\beta \\ -x'_i\beta \end{cases}$$

Cela implique que les  $e_i$  suivent nécessairement une discrète et ne peut pas considérée comme une loi normale.

D'autre part, on a  $E[Y/X] = p(X) \in [0,1]$ , cela impose que  $0 \leq X'\beta \leq 1$ . Les hypothèses de la régression linéaire n'imposent pas de tels critères sur les variables explicatives, par conséquent, on ne peut pas garantir que ce terme soit inclus entre 0 et 1.

En outre, on trouvera de problèmes à interpréter le paramètre  $\beta$  du modèle surtout lorsqu'on recodifie la variable expliquée.

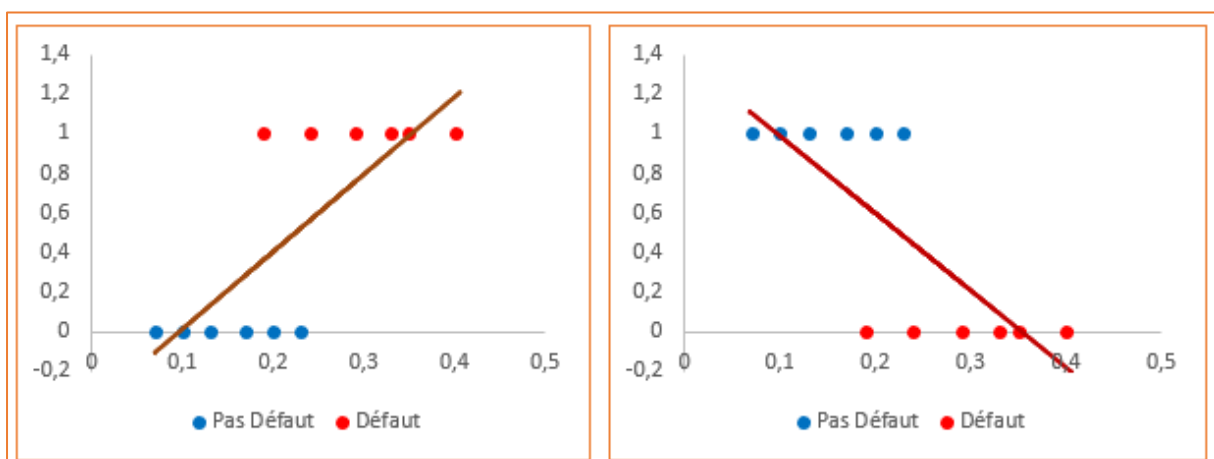


Figure 5 : comparaison des régressions linéaires de la variable défaut en modifiant la codification

Les deux graphes précédents résument les problèmes déjà cités :

- La droite de régression n'est pas la même si on change la codification de la variable défaut ; en d'autres termes, on obtient des valeurs différentes pour le paramètre  $\beta$  du modèle.
- La droite de régression qui doit représenter une probabilité, sort de l'intervalle  $[0,1]$ .

Pour résoudre ces problèmes, on aura besoin d'une transformation qui modélise cette probabilité tout en vérifiant les hypothèses suivantes :

- Bijective et dérivable
- Définie sur l'intervalle  $[0,1]$  vers  $]-\infty, +\infty[$

Cette transformation nous permettra finalement de modéliser les probabilités avec un sigmoïde au lieu d'une droite de régression.

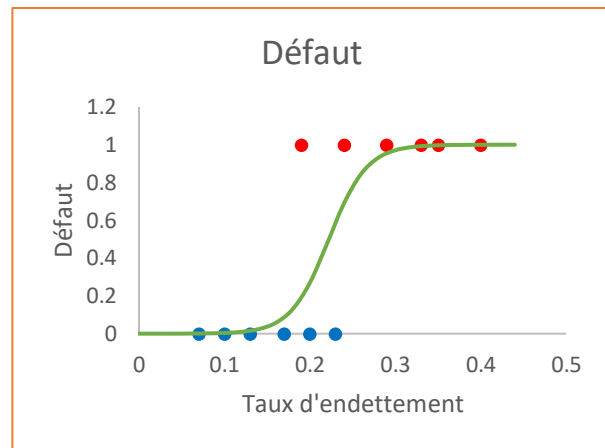


Figure 6 : Modélisation de la probabilité de défaut par un sigmoïde

## 1.2.Odds

Inspiré du domaine de la loterie, lorsqu'on dit que les chances (Odds) pour que mon équipe gagne est 5 sur 3, c'est-à-dire que sur 8 matches, mon équipe peut gagner 5 matches et perdre 3.

Cet Odds peut être écrite sous forme d'un ratio :  $\frac{5}{3} = 1,7$ , bien évidemment, ce ratio n'est pas une probabilité, en revanche, on peut l'écrire en fonction de probabilité en divisant le numérateur et le dénominateur par le nombre total des matches, on aura donc la formule suivante :  $Odds = \frac{\text{probabilité de gagner}}{\text{probabilité de perdre}} = \frac{\text{probabilité de gagner}}{1 - \text{probabilité de gagner}} = \frac{p}{1-p}$

Pour adapter ce ratio à notre étude, on considère que  $p = p(X)$  est la probabilité de défaut.

Ce ratio représente une limite au niveau de la comparaison. En effet ; un Odds de défaut de 5 sur 1 est égal à 5, par contre, un Odds de 1 sur 5 est égal à 0,2, donc il est difficile de comparer ces deux valeurs.

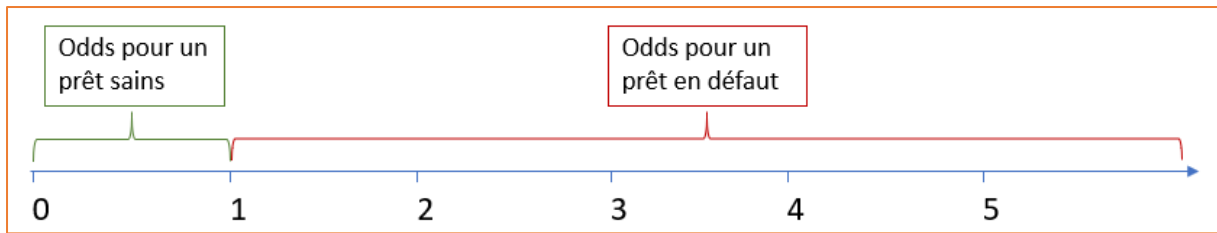


Figure 7 : Asymétrie des Odds

Cette droite numérique nous montre l'asymétrie entre l'intervalle  $[1, +\infty[$  correspond aux valeurs des Odds d'avoir défaut et l'intervalle  $[0,1]$  correspond aux valeurs des Odds de ne pas avoir défaut.

Pour résoudre ce problème, on pense à une transformation logarithmique, car cette fonction transforme l'intervalle  $[0,1]$  vers  $] -\infty, 0]$ , et l'intervalle  $[1, +\infty[$  vers  $[0, +\infty[$ . Donc pour l'exemple précédent, on aura  $\log\left(\frac{1}{5}\right) = -\log\left(\frac{5}{1}\right) = -1,609$ . La comparaison dans ce cas devient triviale, les deux valeurs sont séparées par la même distance à l'origine.

En simulant cette fonction, appelée logit, sur 1000 probabilités générées aléatoirement, on trouve la distribution suivante :

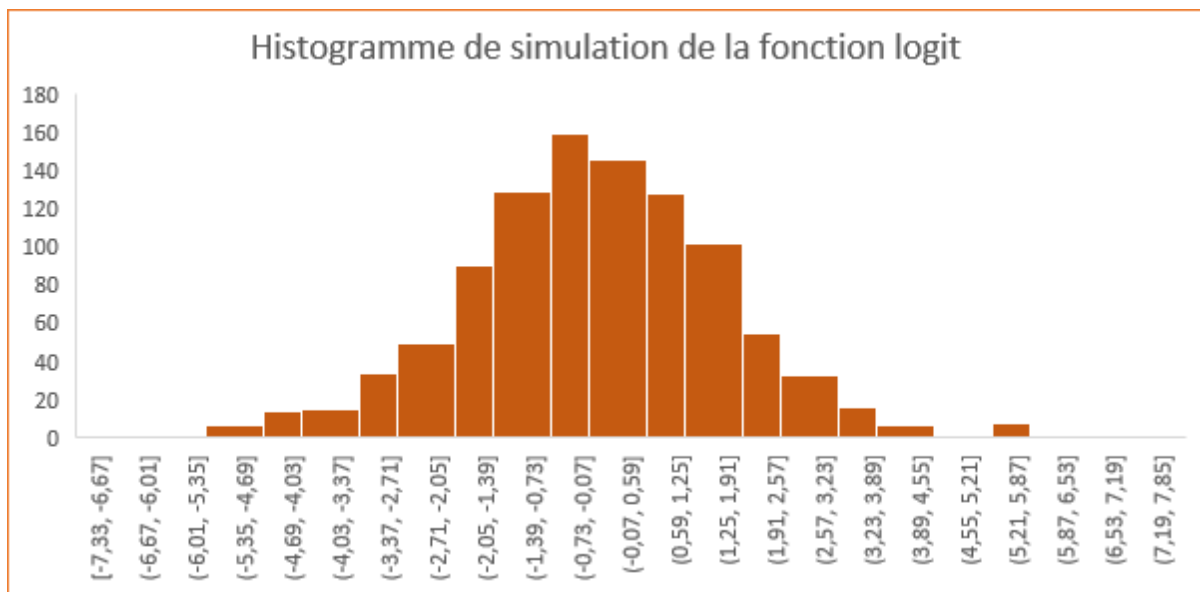


Figure 8 : Distribution de la fonction logit : simulation sur 1000 probabilités

D'après cet histogramme, on constate que la distribution de l'image de la fonction logit est semblable à une loi normale.

En revenant à notre problématique de départ, la fonction  $logit = \log\left(\frac{p}{1-p}\right)$  vérifie les hypothèses de la transformation qu'on cherche :

- La fonction logit est bijective et dérivable
- Elle permet de transformer l'intervalle  $[0,1]$  vers  $\mathbb{R}$ .
- La distribution de l'image de la fonction suit une loi normale.

Pour ces raisons, au lieu de modéliser la probabilité de défaut  $p(X)$ , on modélisera la fonction logit de cette probabilité, on trouve le modèle suivant :  $logit(p(X)) = logit(E[Y/X]) = \log\left(\frac{p(X)}{1-p(X)}\right) = X'\beta$ .

Inversement, pour trouver les probabilités de défaut, on utilise la fonction inverse de la fonction Logit qu'on nomme F :  $p(X) = F(X'\beta) = \frac{e^{X'\beta}}{1+e^{X'\beta}}$

### Estimation des paramètres

La transformation Logit nous permet d'avoir une nouvelle variable quantitative définie sur  $\mathbb{R}$  avec :  $logit(0) = -\infty$  et  $logit(1) = +\infty$ . Par conséquent, on retrouve un modèle de régression linéaire simple, donc le but serait d'ajuster les données par une droite et estimer ses paramètres.

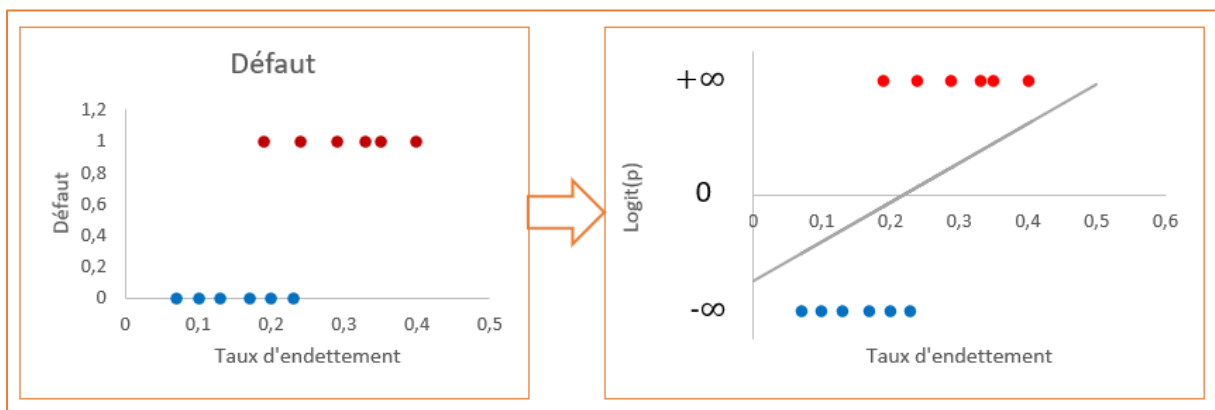


Figure 9 : Transformation de la variable défaut par la fonction logit

L'estimation des paramètres d'un modèle de régression linéaire simple par la méthode des MCO consiste à minimiser la somme des carrés des résidus. Cette méthode ne fonctionne pas dans ce cas, car les observations réelles se situent à l'infini, donc pour toute droite choisie, l'écart entre les valeurs observées et les valeurs estimées serait toujours égale à l'infini.

Pour cela, on choisit d'utiliser la méthode d'estimation par le maximum de vraisemblance.

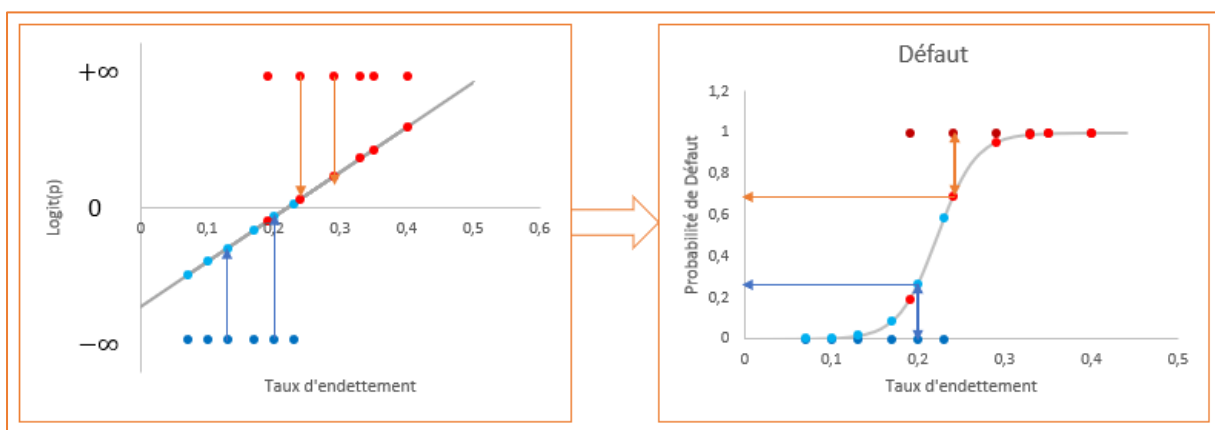


Figure 10 : Principe d'utilisation de la vraisemblance

Comme le montre la figure 7, après la transformation des données par la fonction Logit, on choisit une droite, puis on projette les observations sur cette droite pour avoir des valeurs de la variable Logit(p). Ensuite on calcule les probabilités correspondantes par la fonction inverse du Logit, qui vont constituer le sigmoïde (l'image à droite). Finalement, on calcule la fonction de vraisemblance entre ces probabilités et les probabilités réelles. On répète le processus jusqu'à trouver la droite où cette fonction est maximale.

### 1.3.Fonction de vraisemblance

La fonction de vraisemblance permettant l'estimation des paramètres  $\beta$  du modèle s'écrit sous la forme :  $L(\beta) = \prod_{i=1}^n p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}$

Avec  $\begin{cases} n \text{ désigne le nombre d'individus} \\ y_i : \text{la probabilité réelle de l'individu } i \text{ (1 en cas de défaut, 0 sinon)} \\ p(x_i) : \text{la probabilité de défaut pour l'individu } i \end{cases}$

Pour simplifier les calculs, on utilise le logarithme de la fonction de vraisemblance :

$$l(\beta) = \ln(L(\beta)) = \sum_{i=1}^n y_i \ln(p(x_i)) + (1 - y_i) \ln(1 - p(x_i))$$

Ce qui implique :  $l(\beta) = \sum_{i=1}^n y_i x'_i \beta - \ln(1 - \exp(x'_i \beta))$ , cette fonction est globalement concave, donc le gradient nul correspond au son maximum.

En conséquence, les équations à résoudre sont :  $\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^n x_{ij} (y_j - p(x_i)) = 0$

Ces équations ne sont pas linéaires ;  $p(X) = F(X'\beta)$ , donc on ne peut pas avoir un estimateur explicite de  $\beta$ . La famille exponentielle à laquelle la loi de Bernoulli appartient, nous garantit une solution unique pour ces équations. La résolution se fait à l'aide des méthodes numériques comme par exemple Newton-Raphson.

### 1.4.Propriétés des estimateurs

Les estimateurs du maximum de vraisemblance convergent en loi vers la loi normale :

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow \mathfrak{N}(\beta, I^{-1}(\hat{\beta}))$$

Avec  $I^{-1}(\hat{\beta})$  est l'inverse de la matrice d'information de Fisher

Ce résultat nous permet de construire approximativement un intervalle de confiance au niveau de confiance  $1 - \alpha$  pour chaque paramètre  $\beta_j$  :  $\left[ \hat{\beta} - z_{\frac{\alpha}{2}} \frac{\hat{\sigma}_j}{\sqrt{n}} ; \hat{\beta} + z_{\frac{\alpha}{2}} \frac{\hat{\sigma}_j}{\sqrt{n}} \right]$

Avec  $\hat{\sigma}_j$  représente le j<sup>ème</sup> terme de la diagonale de la matrice d'information  $I^{-1}(\hat{\beta})$

Pour tester la significativité de ces paramètres estimés, autrement dit, évaluer la contribution des variables, on utilise le test de Wald.

Formellement, les hypothèses du test s'écrivent de la manière suivante : 
$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

En s'appuyant sur la normalité asymptotique des estimateurs de vraisemblance, la statistique du test s'écrit sous la forme : 
$$W_j = \frac{\widehat{\beta}_j^2}{\widehat{\sigma}_j^2}$$

Sous l'hypothèse nulle  $H_0$ , cette statistique suit une loi de  $\chi^2$  à 1 degré de liberté. Si la valeur observée  $W_{obs} > \chi_{1-\alpha}^2$ , on rejette  $H_0$  au seuil  $\alpha$  ; Dans ce cas, la variable  $X_j$  a un effet sur la probabilité - de défaut dans notre étude - sachant les autres variables.

### 1.5.Méthode de sélection des variables

Le modèle parfois peut contenir certaines variables qui n'apportent pas d'informations dans la modélisation, Pour cela, il existe 3 méthodes qui permettent de sélectionner que les variables les plus importantes, tout en optimisant un critère d'information tel que l'AIC ou le BIC :

- Méthode Forward : à chaque fois, une variable est ajoutée au modèle, c'est celle qui permet de réduire au mieux le critère AIC du modèle obtenu. La procédure s'arrête lorsque toutes les variables sont introduites ou lorsque l'AIC ne décroît plus.
- Méthode Backward : l'algorithme démarre cette fois du modèle complet. A chaque étape, la variable dont l'élimination conduit à l'AIC le plus élevé est éliminée. La procédure s'arrête lorsque l'AIC ne décroît plus.
- Méthode Stepwise : c'est une méthode mixte entre les deux autres méthodes, l'algorithme démarre également du modèle complet, après élimination d'une variable comme dans le cas de la méthode Backward, on ajoute les autres variables éliminées. Le modèle optimal est celui dont l'AIC est le plus faible.

## 2. Arbre de classification

### 2.1. Définition

L'arbre de classification est une méthode d'apprentissage automatique sous forme d'arbre de décision. Cet arbre modélise une hiérarchie de tests (questions sur les variables dont les réponses sont "vrai" ou "faux" pour les arbres binaires). Il permet finalement de prédire la modalité ou la classe des individus dans le cas d'une variable qualitative telle que le défaut.

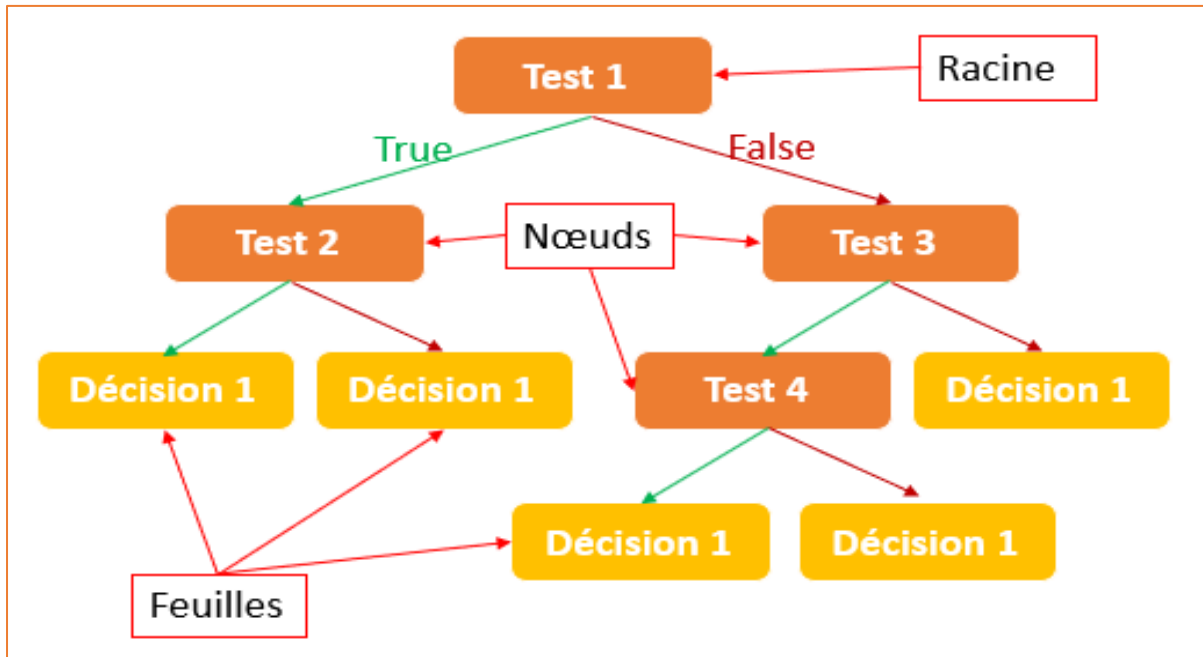


Figure 11 : Composantes d'un arbre de classification

Un arbre de classification se compose principalement de 3 éléments :

- Les nœuds : on distingue entre de types de nœuds :
  - La racine : le premier sommet qui concerne le premier test, dont il y a que des flèches sortantes.
  - Des nœuds intermédiaires : des nœuds des niveaux inférieurs apportant des tests supplémentaires pour partitionner encore les individus, pour ces nœuds, on constate des flèches entrantes et des flèches sortantes.

Les critères et méthodes de détermination des tests et les variables pour chaque nœud (la racine et les nœuds intermédiaires) sont les mêmes.

- Les flèches : elles indiquent le sens du chemin à suivre jusqu'à la décision finale. Dans notre cas on utilise des arbres binaires, c'est-à-dire, de chaque nœud, on aura deux flèches sortantes ; une indique la décision à prendre ou le test suivant si la condition dans le nœud est vraie (les flèches vertes dans la figure 8), et l'autre dans le cas contraire (les flèches rouges dans la figure 8).

- Les feuilles : sont les nœuds des derniers niveaux de l'arbre, elles contiennent la décision ou la classe prédite pour l'individu. Evidemment, il n'y a que des flèches entrantes pour ces feuilles.

Dans un problème de classification, la décision au niveau de cette feuille est prise par rapport à la modalité qui représente le plus grand nombre d'individus, si tous les individus dans cette feuille ont la même modalité, elle est considérée comme pure.

## 2.2. Construction d'un arbre de classification

Pour construire un arbre de classification, la première question qui se pose : quelle variable choisir en premier, ensuite on pose la question de quelle est la condition à tester sur cette variable ; il est clair que pour une variable explicative dichotomique, le partitionnement serait de tester si cet individu vérifie l'une des deux modalités, donc quel serait le cas pour une variable ayant plus de modalités, ou bien pour des variables continues ?

Pour choisir la variable qui réalise la meilleure séparation d'individus, on utilisera un critère de segmentation qu'on calcule pour les différentes variables d'apprentissage (explicatives). Par conséquent, la variable choisie serait celle qui optimise ce critère donné. Le critère le plus utilisé pour les arbres de classification est le critère de diversité de Gini.

En général, l'indice de Gini est un indicateur qui mesure le niveau d'inégalité de la répartition d'une variable (tel que le revenu) sur une population donnée.

L'indice de diversité de Gini utilisé par l'algorithme CART, mesure la fréquence pour qu'un individu serait mal classé par le test du nœud.

$$I_G(f) = \sum_{i=1}^m f_i(1 - f_i) = 1 - \sum_{i=1}^m f_i^2 \text{ Avec :}$$

- $m$  Représente le nombre de modalités de la variable à expliquer. Dans notre étude, il s'agit d'une variable dichotomique, le nombre de ses modalités est égale à 2 : "Défaut" ou "Pas défaut".
- $f_i$  : représente la fréquence ou la probabilité de la modalité  $i$  dans la feuille après la répartition, c'est-à-dire, le ratio du nombre d'individus qui ont cette modalité dans cette feuille, sur le total des individus dans cette feuille.

L'adaptation de la formule précédente à notre variable dépendante, nous permet d'écrire une formule plus simple :  $I_G = 1 - (\text{proba}(\text{Défaut}))^2 - (\text{proba}(\text{Pas Défaut}))^2$

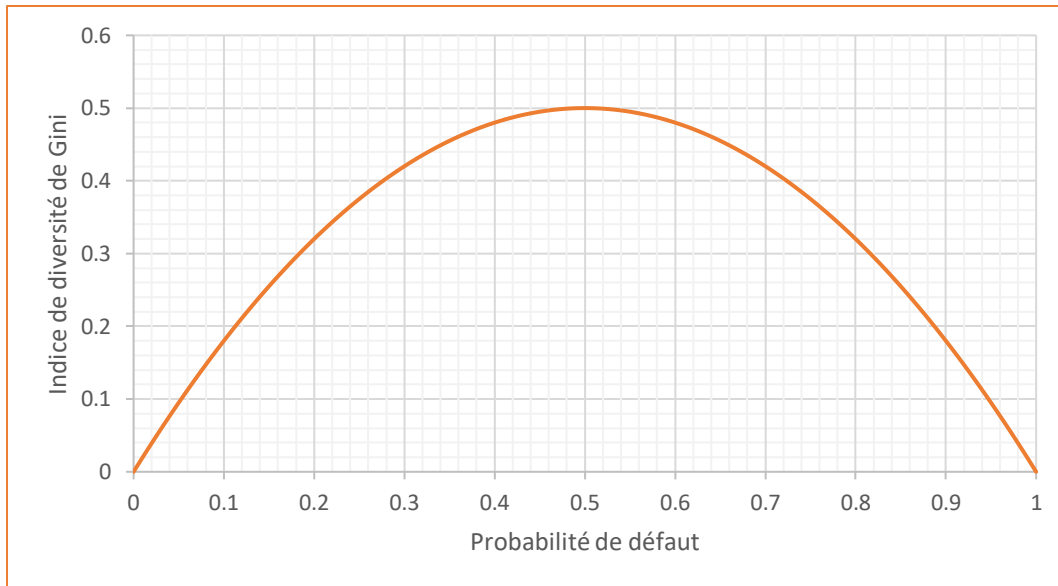


Figure 12 : Variation de l'indice de diversité de Gini

Cet indice prend des valeurs entre 0 et 0,5 comme le montre le graphe suivant, il atteint le 0 lorsque la probabilité de défaut est égale à 0 ou à 1. Autrement dit, si tous les individus dans cette feuille sont tous en défaut, ou sont tous sains. Idéalement, on cherche à avoir l'un des deux cas dans la feuille pour que la répartition soit parfaite. Dans ce cas, la feuille est considérée pure.

En conclusion, le choix de la meilleure répartition consiste à minimiser cet indice de Gini.

### 2.3.Exemple

On suppose que la variable "Sexe" est parmi les variables explicatives, elle prend 2 valeurs : "M" pour un homme et "F" pour une femme.

La condition à tester serait si cette variable est égale une de ces deux modalités, on prend par exemple le sexe masculin.

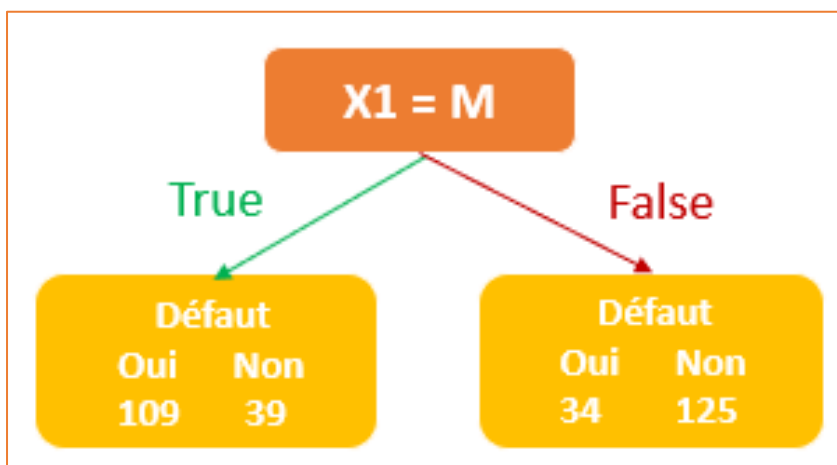


Figure 13 : Exemple pour la construction d'un arbre de classification

On répartit les individus de notre base d'apprentissage (307 dans cet exemple) sur les deux feuilles en indiquant le nombre de "Défaut" et le nombre des crédits sains.

On calcule l'indice de diversité de Gini pour chaque feuille :

$$I_G(i) = 1 - \left( \frac{\text{nombre de défauts dans la feuille } i}{\text{nombre total d'individus dans la feuille } i} \right)^2 - \left( \frac{\text{nombre de crédits sains dans la feuille } i}{\text{nombre total d'individus dans la feuille } i} \right)^2$$

- Feuille à gauche :  $I_G(1) = 1 - \left( \frac{109}{109+39} \right)^2 - \left( \frac{39}{109+39} \right)^2 = 0,388$
- Feuille à droite :  $I_G(2) = 1 - \left( \frac{34}{125+34} \right)^2 - \left( \frac{125}{125+34} \right)^2 = 0,336$

Après avoir calculé les indices de diversité pour chaque feuille, on aura besoin de calculer un indice global pour cette répartition pour qu'elle soit comparable avec d'autres répartitions. Pour ce faire, l'indice global est égal à la moyenne pondérée des indices de Gini de chaque

$$\text{feuille : } I_G(\text{Global}) = \sum_{i=1}^2 \frac{\text{nombre d'individus dans la feuille } i}{\text{nombre total d'individus dans la répartition}} \times I_G(i)$$

Donc l'indice globale de diversité de Gini pour la répartition de l'exemple précédent est :

$$I_G(\text{Global}) = \frac{148}{307} \times 0,388 + \frac{159}{307} \times 0,336 = 0,361$$

Ensuite, on calcule cet indice pour les autres variables, et choisit la variable dont l'indice est minimal.

On peut également calculer un indice en cas d'aucune répartition :

$$I_G(\text{Aucune répartition}) = 1 - \left( \frac{109+34}{307} \right)^2 - \left( \frac{39+125}{307} \right)^2 = 0,497$$

Donc la segmentation par la variable X1 est plus pertinente que ne rien faire.

Après avoir vu comment choisir les variables, on verra par la suite comment choisir le test ou bien la condition à vérifier dans le nœud pour les différents types de variables :

- Variable binaire : comme on vient de voir dans l'exemple précédent, la condition serait de tester si l'individu vérifie l'une des deux modalités.
- Variable numérique à valeur dans  $\mathbb{R}$  : pour ce type de variables, l'algorithme trie les individus par ordre croissant selon cette variable, ensuite il calcule les moyennes entre deux individus consécutifs, finalement il calcule l'indice globale de diversité de Gini en mettant comme test dans le nœud : "Est-ce que la valeur de la variable en question est inférieure à cette moyenne ?". Il refait la même démarche jusqu'à trouver le seuil qui minimise l'indice de Gini.

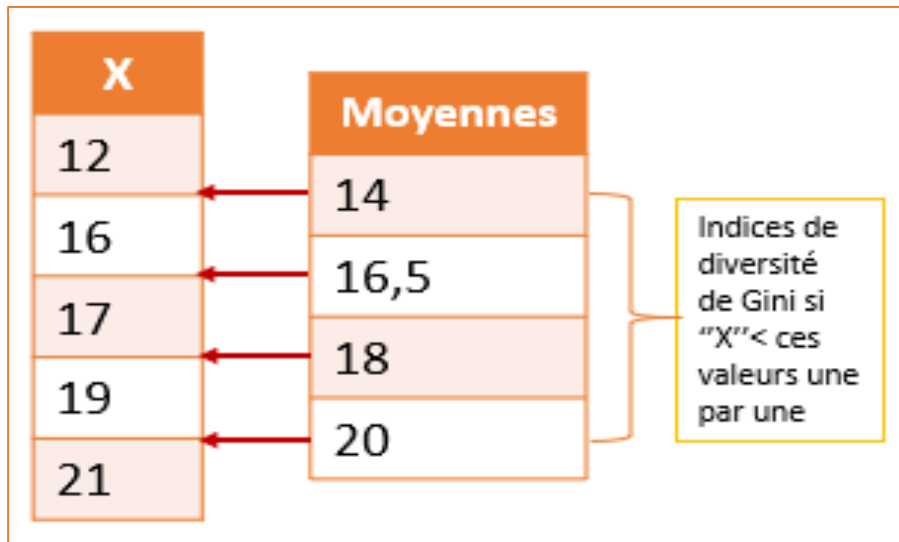


Figure 14 : Calcul indice de Gini pour une variable qualitative

- Variable qualitative ordinale : on les traite de presque de la même façon que dans le cas de variables numériques sans calculer des moyennes, à chaque fois on calcule l'indice de Gini en testant si la variable est inférieure ou égale à un rang.
- Variable qualitative nominale : on calcule l'indice de Gini pour toutes les combinaisons possibles.

## 2.4.Problème de surajustement

‘les arbres ont un aspect qui les empêche d’être l’outil idéal pour l’apprentissage, nommé *inaccuracy (inexactitude)*’<sup>9</sup>, cette inaccuracy concerne les prédictions sur des observations qui n’existent pas dans la base d’apprentissage, autrement dit, ces arbres construisent de bons modèles pour les données sur lesquelles elles étaient créées, mais elles ne sont pas flexibles pour la classification d’un nouvel échantillon.

En outre, l’algorithme peut segmenter les variables jusqu’à trouver un nombre très réduit dans une ou plusieurs feuille (une feuille contenant 5 individus sur un échantillon de 300 individus par exemple : 4 en défaut et le 5<sup>ième</sup> sain). Ce problème diminue l’exactitude dans la prédiction pour un nouvel échantillon puisque l’effectif n’est pas assez suffisant pour conclure la décision convenable.

<sup>9</sup> Traduite de l’anglais : ‘Trees have one aspect that prevent them from being the ideal tool for learning, namely inaccuracy’ du livre The Element of Statistical Learning, J.Friedman.

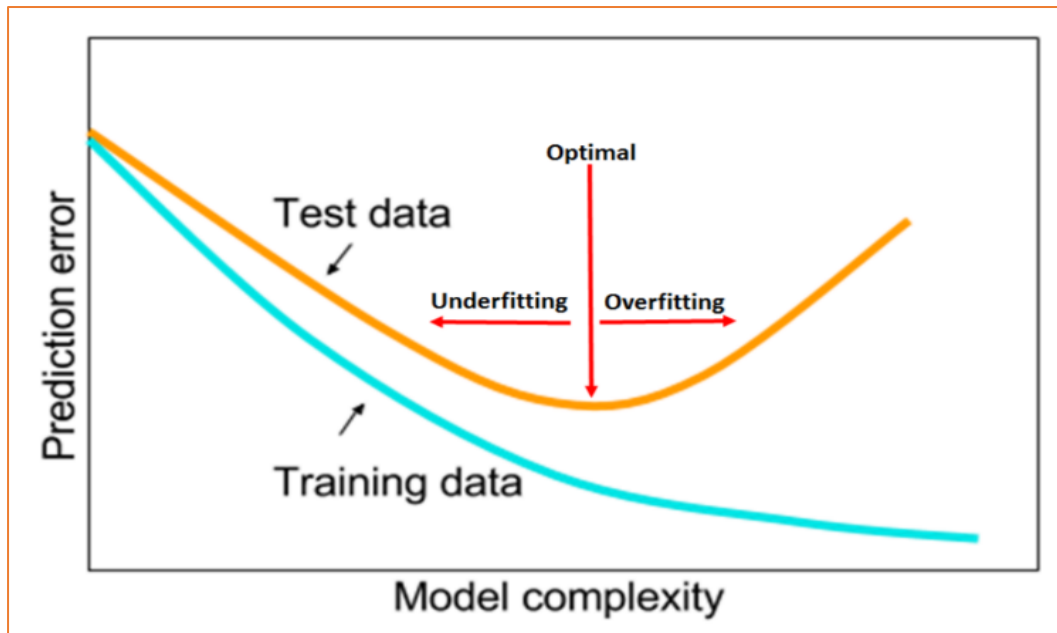


Figure 15<sup>10</sup> : Problème de surajustement

La figure 12 nous montre que si la complexité du modèle augmente, c'est-à-dire le nombre d'itérations ou bien le nombre de feuilles dans l'arbre, l'erreur de prédiction sur la base d'apprentissage décroît. Ce résultat semble trivial ; dans le cas extrême, où le nombre de feuilles est égale au nombre d'individus, on aura un taux d'erreur de prédiction nul, donc plus le nombre de feuilles augmente, plus l'effectif dans ces feuilles diminue et par conséquent le taux d'erreur de prédiction diminue. En contrepartie, l'erreur de prédiction sur une base de test, diminue avec l'augmentation du nombre d'itérations, et à partir d'un certain seuil, cette erreur commence à augmenter, cela est dû aux effectifs insuffisants dans les feuilles. Ce phénomène s'appelle le surajustement (overfitting).

Le problème de surajustement apparaît lorsque le modèle ajuste parfaitement la base d'apprentissage, alors que la prédiction sur des nouvelles observations est faible.

Pour résoudre ce problème, on procède par des méthodes d'élagage pour choisir la taille optimale de l'arbre, on distingue principalement entre deux méthodes :

- Le pré-élagage : cette méthode consiste à proposer des conditions d'arrêt lors de la phase d'expansion. Autrement dit, des critères pour arrêter la séparation de l'échantillon avant la construction de la base, comme par exemple un effectif minimal dans la feuille, un effectif minimal pour la séparation, ou l'homogénéité des individus dans la feuille ...

<sup>10</sup> Source : <https://towardsdatascience.com/hyper-parameter-tuning-techniques-in-deep-learning-4dad592c63c8>  
date de consultation : Aout 2020

- Le post-élagage : cette méthode fonctionne sur deux phases ; la première consiste à construire l'arbre dans sa taille maximale sur une base d'apprentissage, la deuxième phase consiste à réduire le nombre de ses feuilles afin de minimiser le taux d'erreur de prédiction sur une base de test.

### 3. Forêt aléatoire

La forêt aléatoire (Random Forest) est une technique d'apprentissage automatique, elle effectue l'apprentissage sur plusieurs arbres entraînés sur des sous ensemble de données.

Afin de créer un grand nombre d'arbres, on utilise la technique du bootstrap.

#### 3.1. Bootstrap

Le bootstrap est une méthode qui nous permet de créer des échantillons à partir d'un jeu de données, elle est basée sur la réplication multiple des données. Autrement dit, cette méthode nous permet de construire une base en tirant des observations d'une façon aléatoire (avec répétition) à partir de la base initiale.

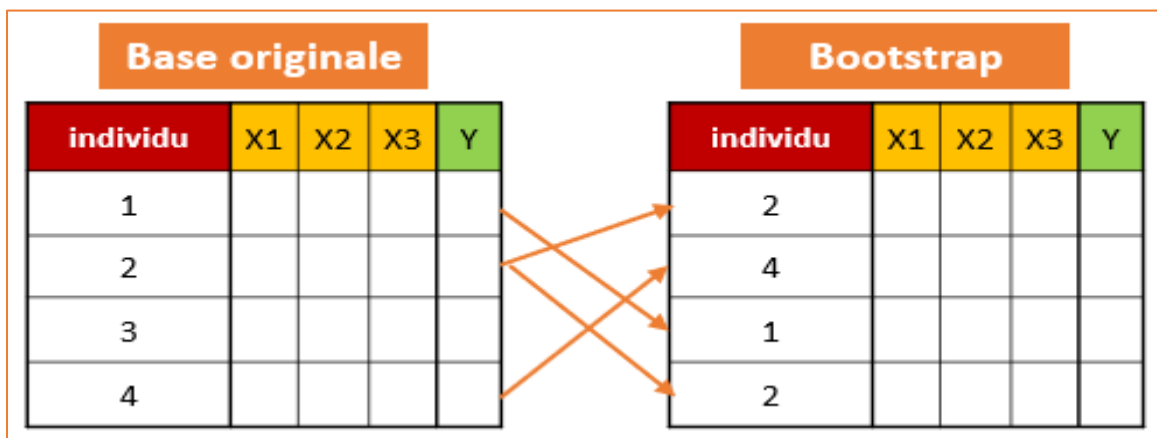


Figure 16 : Principe du Bootstrap

Comme le montre la figure, l'échantillon créé par le Bootstrap peut contenir des observations répétées, et bien évidemment si la taille de l'échantillon créé par le Bootstrap est la même que pour la base originale, certains individus n'apparaissent pas dans cet échantillon. Les travaux montrent qu'environ d'un tiers de la base ne figure pas dans cet échantillon.

Dans le but d'estimer des paramètres ou la construction d'un modèle, le principe du Bootstrap consiste à créer plusieurs échantillons aléatoires et identiquement distribués, puis agréger les résultats obtenus. Cette méthode (Bootstrap + Agrégation) s'appelle Bagging

#### 3.2. Construction d'une forêt aléatoire

Grace aux techniques de Bootstrap, on arrive à créer plusieurs échantillons aléatoires à partir de la base de données. L'exploitation de ces échantillons nous permet de construire plusieurs arbres de classification différentes, ce qui est équivalent à une "forêt", d'où le nom du modèle

**Algorithm 15.1** *Random Forest for Regression or Classification.*

1. For  $b = 1$  to  $B$ :
  - (a) Draw a bootstrap sample  $\mathbf{Z}^*$  of size  $N$  from the training data.
  - (b) Grow a random-forest tree  $T_b$  to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size  $n_{min}$  is reached.
    - i. Select  $m$  variables at random from the  $p$  variables.
    - ii. Pick the best variable/split-point among the  $m$ .
    - iii. Split the node into two daughter nodes.
2. Output the ensemble of trees  $\{T_b\}_1^B$ .

To make a prediction at a new point  $x$ :

*Regression:*  $\hat{f}_{\text{rf}}^B(x) = \frac{1}{B} \sum_{b=1}^B T_b(x)$ .

*Classification:* Let  $\hat{C}_b(x)$  be the class prediction of the  $b$ th random-forest tree. Then  $\hat{C}_{\text{rf}}^B(x) = \text{majority vote } \{\hat{C}_b(x)\}_1^B$ .

Figure 17<sup>11</sup> : Algorithme de construction d'une forêt aléatoire

La création des arbres par Random Forest diffère de ce qu'on a vu précédemment ; pour créer un arbre individuel, on sélectionne la répartition dont l'indice de diversité de Gini est minimal en comparant toutes les variables, la spécificité des Random Forests réside dans l'utilisation d'un nombre réduit de variables tirés aléatoirement pour le choix des meilleures répartitions dans ses arbres.

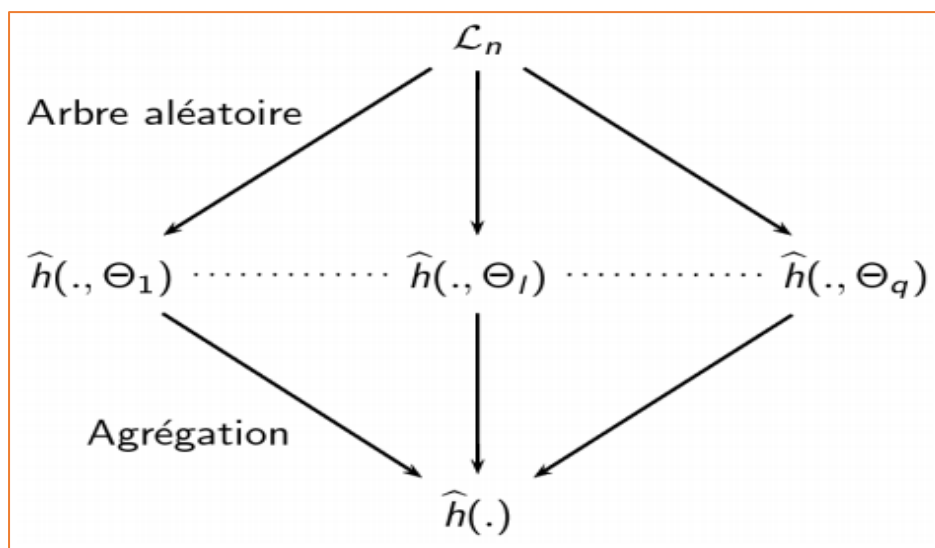


Figure 18 : Schéma du processus du Random Forest

<sup>11</sup> The Element of Statistical Learning, J.Friedman.

### 3.3. Out Of Bag

Lors de la création d'un échantillon par la méthode du Boosttrap, on autorise la répétition de dans la sélection des individus, en conséquence, certains individus ne participent pas à la constitution de cet échantillon, ces individus sont appelés Out-Of-Bag.

Puisque l'OBB ne participe pas à la création de l'arbre, on peut l'utiliser comme un échantillon de test pour mesurer la performance de la classification ; en effet, on exécute l'OBB sur tous les arbres créés sans cet OOB, on l'attribut la modalité avec le maximum des votes par ces arbres, et on la compare avec la valeur observée. On refait cela pour tous les OOB dans le but de construire une table de confusion qui nous permet par la suite de mesurer l'erreur de prédiction par cette Random Forest. La proportion d'individus OOB qui étaient mal classifiés s'appelle l'erreur d'Out-Of-Bag (OOB error).

### 3.4. Prédiction

Pour prédire la valeur d'une observation à l'aide d'un modèle de Random Forest, l'algorithme l'exécute sur tous les arbres du modèle. La valeur prédite serait la classe qui a eu plus de vote par ces arbres. Dans le cas d'un problème de régression, la valeur prédite serait la moyenne sur ces arbres.

Les Random Forêts permettent d'améliorer la performance de prédiction par rapport à un arbre individuel. Par contre, elles ne sont pas interprétables.

## 4. Gradient boosting

### 4.1. Définition

Le Boosting est une technique d'agrégation de modèles permettant d'obtenir un estimateur fort à partir d'un ensemble d'estimateurs faibles (weak learners ou base learners). Les méthodes de Boosting utilisent généralement les arbres de décision comme des estimateurs de base, dans ce cas, l'objectif de ces méthodes consiste à corriger le problème de surajustement des arbres pour améliorer la qualité des prédictions.

Le principe d'agrégation de modèles nous rappelle des méthodes de Bagging, et plus précisément les forêts aléatoires puisqu'elles sont les arbres. La différence entre ces deux méthodes réside dans la construction des arbres. Dans le cas des forêts aléatoires, les arbres sont créés d'une façon indépendante à partir des échantillons tirés aléatoirement de la base d'apprentissage, les pertitions sont une forme de moyennes sur ces arbres dans le cas de régression, et la classe qui a eu le plus grand nombre de vote pour la classification. Par contre, les méthodes de Boosting créent les arbres d'une manière récursive : chaque arbre est créé par la régression des erreurs commises par l'arbre précédente.

### 4.2. Arbre de régression CART

La construction d'un arbre de régression est similaire à la construction d'un arbre de classification, les différences entre les deux modèles concernent les prédictions et les critères de séparation.

La valeur prédite  $\hat{c}_j$  par une feuille  $R_j$  n'est qu'une moyenne de la variable expliquée de tous les individus appartenant à cette feuille :  $\hat{c}_j = \text{moyenne}(y_i | x_i \in R_j)$

Afin de choisir la meilleure répartition dans un nœud, qui divise les observations en deux régions :  $R_1$  et  $R_2$ , le critère le plus souvent utilisé est la somme des carrés des résidus entre les valeurs observées et les valeurs prédites (les moyennes des classes).

On considère  $s$  le seuil de séparation pour la variable  $X_j$ , les deux régions seront définies comme suit :  $R_1(j, s) = \{X | X_j \leq s\}$  et  $R_2(j, s) = \{X | X_j > s\}$ , ensuite on cherche la variable  $j$  et le seuil  $s$  qui vérifient : 
$$\min_{j,s} \left[ \min_{c_1} [\sum_{x_i \in R_1(j,s)} (y_i - c_1)^2] + \min_{c_2} [\sum_{x_i \in R_2(j,s)} (y_i - c_2)^2] \right].$$

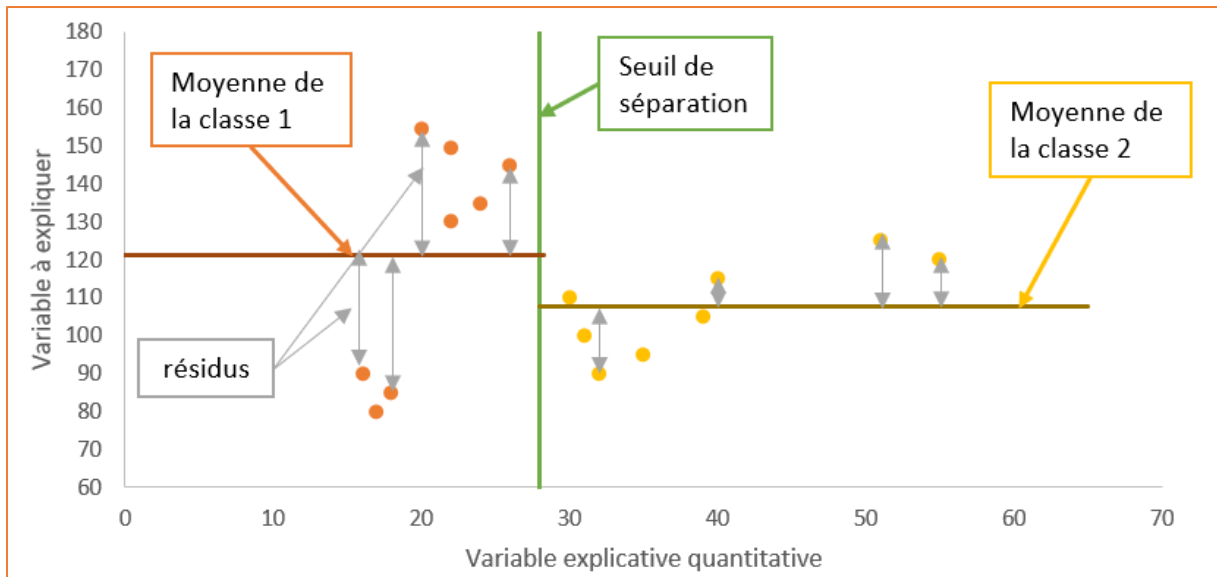


Figure 19 : Arbre de régression : Choix du seuil de séparation pour une variable quantitative

Pour une variable explicative quantitative, on varie le seuil de séparation, qui est la moyenne de cette variable pour deux observations successives, jusqu'à trouver la valeur qui minimise la somme des carrés des résidus.

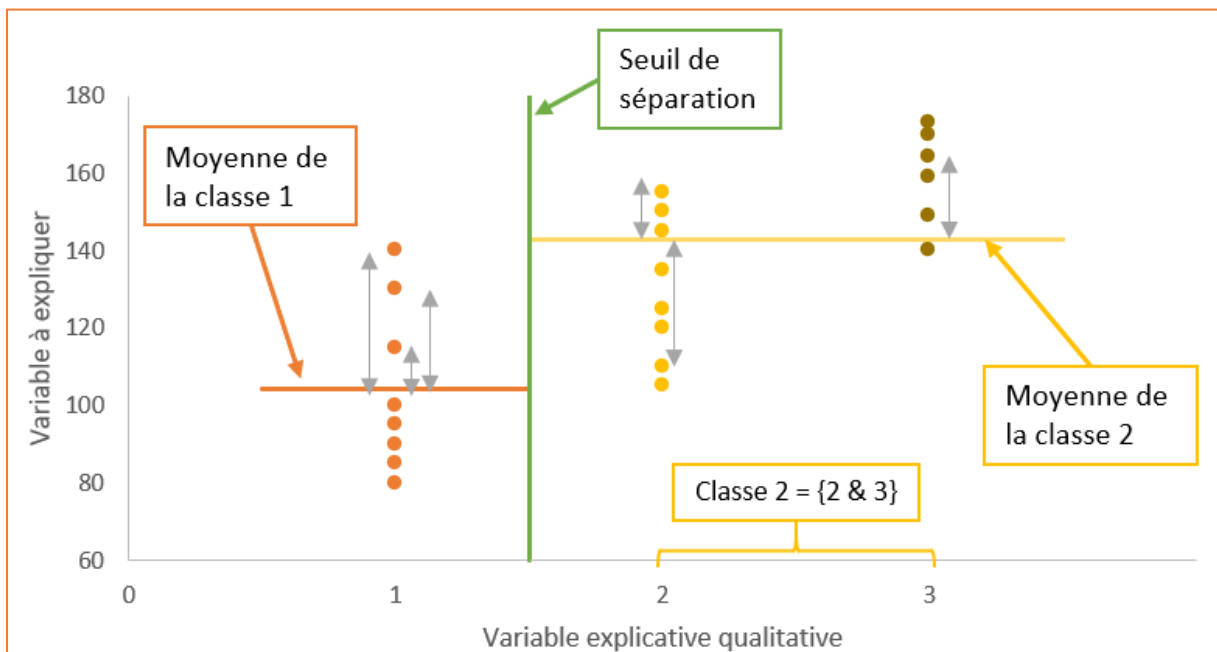


Figure 20 : Arbre de régression : Choix de la meilleure combinaison de séparation pour une variable qualitative

Dans le cas d'une variable qualitative, on applique la même technique sur les différentes combinaisons possibles de ses modalités.

Toutes les autres propriétés des arbres de classification, resteront valable également pour ces arbres.

### 4.3. Fonction de perte

Une fonction de perte mesure la pénalité d'une mauvaise prédiction. En d'autres termes, la perte correspond à un nombre qui indique la médiocrité de la prédiction d'un modèle. Elle prend sa valeur minimale, zéro, lorsque la prédiction du modèle est parfaite. Le principe d'apprentissage des modèles en Machine Learning consiste à minimiser cette fonction.

Cette fonction est généralement mesurée par la différence entre les valeurs observées et les valeurs prédites. Par exemple, dans le cas d'un problème de régression, parmi les fonctions de perte les plus populaires, on trouve la perte quadratique, c'est-à-dire le carré de la différence entre la valeur observée et la valeur prédite :  $(\text{observée} - \text{prédite}(x))^2$ , la somme de ces fonctions sur toutes les observations nous donne la somme des carrés des résidus qu'on essaye de minimiser dans les modèles de régression linéaires.

Pour un modèle de classification, on peut construire une fonction de perte à partir du logarithme de la vraisemblance utilisée dans la régression logistique.

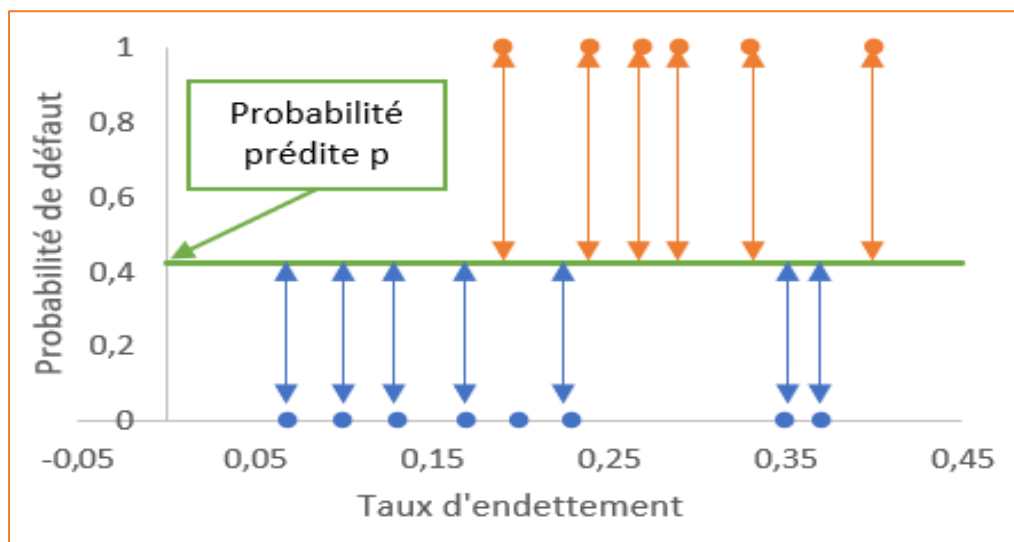


Figure 21 : Fonctionnement d'une fonction de perte

Graphiquement, on peut considérer une fonction de perte comme étant la différence entre les probabilités observées et les probabilités prédites, ou bien une perte quadratique comme dans le cas échéant.

Puisqu'on connaît déjà une fonction à optimiser pour ce type de problèmes, on démarre de ce point pour obtenir une fonction de perte convenable.

Dans la régression logistique, le but était de maximiser le logarithme de la fonction de vraisemblance entre les probabilités prédites et les probabilités pour obtenir les bonnes estimations. On a vu également que cette fonction prend sa valeur maximale, zéro, lorsque le modèle est parfait. Intuitivement, on peut la considérer comme une fonction de perte en la multipliant par  $-1$  pour retrouver un problème de minimisation.

On a  $-l(p) = -\ln(L(p)) = -\sum_{i=1}^n y_i \ln(p(x_i)) + (1 - y_i) \ln(1 - p(x_i))$

La fonction de perte est parfois utile pour un seul individu, comme dans le cas de notre étude.

Par conséquent, cette fonction serait de la forme suivante :  $-[y_i \ln(p(x_i)) + (1 - y_i) \ln(1 - p(x_i))]$ .

On a déjà vu la relation entre la probabilité et le logarithme des Odds, Donc cette fonction peut s'écrire en fonction du  $\log(Odds)$  :  $-y_i \log(Odds(x_i)) + \log(1 + e^{\log(Odds(x_i))})$ .

Finalement la fonction de perte dérivée du logarithme de vraisemblance est :

$$L(y_i; f(x_i)) = -y_i \log(f(x_i)) + \log(1 + e^{\log(f(x_i))})$$

#### 4.4. Algorithme

##### Algorithm 10.3 Gradient Tree Boosting Algorithm.

1. Initialize  $f_0(x) = \arg \min_{\gamma} \sum_{i=1}^N L(y_i, \gamma)$ .

2. For  $m = 1$  to  $M$ :

(a) For  $i = 1, 2, \dots, N$  compute

$$r_{im} = - \left[ \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f=f_{m-1}}.$$

(b) Fit a regression tree to the targets  $r_{im}$  giving terminal regions  $R_{jm}$ ,  $j = 1, 2, \dots, J_m$ .

(c) For  $j = 1, 2, \dots, J_m$  compute

$$\gamma_{jm} = \arg \min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma).$$

(d) Update  $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$ .

3. Output  $\hat{f}(x) = f_M(x)$ .

Figure 22<sup>12</sup> : Algorithme de construction d'un modèle GBM

#### Étape 0 : préparation des données et une fonction de perte

**Input :** Données  $\{(x_i, y_i)\}_{i=1}^N$ , & une fonction de perte dérivable  $L(y_i, f(x_i))$

<sup>12</sup> The Element of Statistical Learning, J.Friedman.

Cette étape consiste à préparer la base d'apprentissage sur laquelle le modèle sera construit ainsi que la fonction de perte à utiliser

Les  $x_i$  représentent les vecteurs des variables explicatives du modèle.

Les  $y_i$  représentent les valeurs observées de la variable à expliquer du modèle.

$N$  représente le nombre d'observations dans la base de d'apprentissage.

$L(y_i, f(x_i))$  représente la fonction de de perte utilisée pour l'évaluation du modèle. La fonction qu'on utilise est  $L(y_i; f(x_i)) = -y_i f(x_i) + \log(1 + e^{f(x_i)})$  avec  $f(x_i)$  représente le  $\log(Odds)$ .

On vérifie que cette fonction est dérivable, avec  $\frac{dL(y_i; f(x))}{df(x)} = -y_i + \frac{e^{f(x)}}{1+e^{f(x)}}$

On a  $f(x) = \log(Odds)$  et  $Odds = \frac{p}{1-p}$  donc  $\frac{e^{f(x)}}{1+e^{f(x)}} = p$

Finalement :  $\frac{dL(y_i; f(x))}{df(x)} = -y_i + p$

Donc la dérivée de la fonction de perte correspond à une sorte de résidu entre la probabilité observée et la probabilité prédite qu'on appelle pseudo résidu.

**Etape 1 : Initialiser  $f_0(x) = \min_{\gamma} \sum_{i=1}^N L(y_i, \gamma)$**

Dans cette étape, on initialise l'algorithme avec une première prédiction constante sur toute la base de d'apprentissage, donc le premier arbre sera sous forme d'une feuille seulement. La valeur de cette feuille  $f_0(x) = \gamma$ , avec  $\gamma$  est la prédiction qui minimise la somme des fonctions de perte.

On a  $\frac{dL(y_i; f(x))}{df(x)} = -y_i + \frac{e^{\gamma}}{1+e^{\gamma}}$  donc  $\frac{d \sum_{i=1}^N L(y_i; f(x))}{df(x)} = \sum_{i=1}^n \left[ -y_i + \frac{e^{\gamma}}{1+e^{\gamma}} \right]$

$\frac{d \sum_{i=1}^N L(y_i; f(x))}{df(x)} = 0$  implique  $\gamma = \log \left( \frac{\sum_{i=1}^N y_i / N}{1 - \sum_{i=1}^N y_i / N} \right)$

Le terme  $\sum_{i=1}^N y_i / N$  représente les observations en défaut sur toutes les observations de l'échantillon, donc il est équivalent à la probabilité de défaut  $p$

D'où  $\gamma = \log \left( \frac{p}{1-p} \right) = \log(Odds)$ , donc la première prédiction du modèle qui minimise la fonction de perte est  $\log(Odds)$ .

**Etape 2 : Pour  $m$  allant de 1 jusqu'à  $M$**

Cette étape est une boucle "pour", qui consiste à construire les  $M$  arbres de régression à partir les variables explicatives dans la base d'apprentissage.

**Etape 2-A : Pour  $i = 1, 2, \dots, N$  Calculer  $r_{i,m} = - \left[ \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f=f_{m-1}}$**

Dans cette partie, on calcule les pseudo-résidus qui correspondent à la différence entre la probabilité observée et la plus récente probabilité prédite dans l'étape ( $m - 1$ ).

En effet, on a  $\frac{dL(y_i, f(x))}{df(x)} = -y_i + p$

Donc  $r_{i,m} = - \left[ \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f=f_{m-1}} = y_i - p(x_i)$

**Etape 2-B : Construire un arbre de régression sur les  $r_{i,m}$  dont les feuilles finales sont  $R_{j,m}, j = 1, 2, \dots, J_m$**

Cette partie consiste à construire un arbre de régression en utilisant les variables explicatives de la base d'apprentissage pour la prédiction des pseudo-résidus.

Ensuite, On recodifie les feuilles de l'arbre  $R_{j,m}, j = 1, 2, \dots, J_m$ , avec  $J_m$  est le nombre de feuilles ou de régions terminale de l'arbre m.

Pour l'instant, cet arbre segmente seulement les individus, on n'a pas encore calculé les valeurs pour la prédiction.

**Etape 2-C : Pour  $j = 1, 2, \dots, J_m$  Calculer  $\gamma_{j,m} = \min_{\gamma} \sum_{x_i \in R_{j,m}} L(y_i, f_{m-1}(x_i) + \gamma)$**

Cette étape est consacrée au calcul des prédictions sur chaque feuille en minimisant la somme des fonctions de pertes des individus appartenant à la feuille.

On a  $\sum_{x_i \in R_{j,m}} L(y_i, f_{m-1}(x_i) + \gamma) = \sum_{x_i \in R_{j,m}} -y[f_{m-1}(x_i) + \gamma] + \log(1 + e^{[f_{m-1}(x_i) + \gamma]})$

Pour minimiser cette somme par rapport à  $\gamma$ , on peut procéder par l'annulation de sa dérivée.

En conséquence, on aura des équations non linéaires pour  $\gamma$ , donc il serait difficile d'obtenir une solution analytique.

Pour remédier à ce problème, on utilisera les approximations polynomiales de Taylor, Dans ce cas, la fonction de perte pour un individus serait :

$$L(y_i, f_{m-1}(x_i) + \gamma) \approx L(y_i, f_{m-1}(x_i)) + \frac{\partial L(y_i, f_{m-1}(x_i))}{\partial f_{m-1}(x_i)} \gamma + \frac{1}{2} \frac{\partial^2 L(y_i, f_{m-1}(x_i))}{\partial f_{m-1}(x_i)^2} \gamma^2$$

En introduisant la somme sur les deux termes :

$$\sum_{x_i \in R_{j,m}} L(y_i, f_{m-1}(x_i) + \gamma) = \sum_{x_i \in R_{j,m}} L(y_i, f_{m-1}(x_i)) + \gamma \times \sum_{x_i \in R_{j,m}} \frac{\partial L(y_i, f_{m-1}(x_i))}{\partial f_{m-1}(x_i)} +$$

$$\frac{1}{2} \gamma^2 \times \sum_{x_i \in R_{j,m}} \frac{\partial^2 L(y_i, f_{m-1}(x_i))}{\partial f_{m-1}(x_i)^2}$$

Cette expression est facile à dériver par rapport à  $\gamma$ , et on obtient donc :

$$\gamma = \frac{-\sum_{x_i \in R_{j,m}} \frac{\partial L(y_i, f_{m-1}(x_i))}{\partial f_{m-1}(x_i)}}{\sum_{x_i \in R_{j,m}} \frac{\partial^2 L(y_i, f_{m-1}(x_i))}{\partial f_{m-1}(x_i)^2}}$$

$$\text{Avec } \frac{\partial L(y_i, f_{m-1}(x_i))}{\partial f_{m-1}(x_i)} = -r_{i,m}$$

$$\text{Et } \frac{\partial^2 L(y_i, f_{m-1}(x_i))}{\partial f_{m-1}(x_i)^2} = \left( \frac{e^{f_{m-1}(x_i)}}{1+e^{f_{m-1}(x_i)}} \right) \frac{1}{1+e^{f_{m-1}(x_i)}} = p(x_i)(1-p(x_i))$$

D'où

$$\gamma = \frac{\sum_{x_i \in R_{j,m}} r_{i,m}}{\sum_{x_i \in R_{j,m}} p(x_i)(1-p(x_i))}$$

**Etape 2-D : Actualiser  $f_m(x) = f_{m-1}(x) + \nu \sum_{j=1}^m \gamma_{j,m} \mathbf{1}_{\{x \in R_{j,m}\}}$**

Dans cette partie, on actualise la prédiction, en ajoutant à la prédiction du rang m-1, l'amélioration de l'arbre construit dans ce rang multiplié par un taux d'apprentissage (Learning Rate)  $\nu$ , ce taux est compris entre 0 et 1. Il représente le poids attribué à l'arbre construit, puisque ce dernier est considéré comme un faible estimateur, il est préférable d'attribuer un poids faible. Par ailleurs, lorsque ce poids est faible, on a intérêt à augmenter le nombre d'arbres.

$\sum_{j=1}^m \gamma_{j,m} \mathbf{1}_{\{x \in R_{j,m}\}}$  Cette somme concerne les individus qui se trouvent dans plus qu'une feuille (région terminale).

**Etape 3 : Output  $\hat{f}(x) = f_M(x)$**

Cette étape concerne la prédiction finale.

## 5. Comparaisons de la qualité des modèles

### 5.1. Séparation de bases : Apprentissage / test

Pour mesurer la qualité d'ajustement d'un modèle, il est intéressant de mesurer sa performance prédictive sur des nouvelles données. Pour cela on divise la base en deux sous-bases, une réservée à l'apprentissage et la construction du modèle, dans ce mémoire, elle constitue 75% de la base totale. Et l'autre réservée au test. Cette méthode nous permet d'évaluer le modèle sur des données nouvelles.

### 5.2. Cross Validation

La séparation de la base en deux parties : base d'apprentissage et base de test, permet d'entraîner le modèle sur la première et de tester sa performance sur la deuxième.

Cette méthode a pour but le test de la performance du modèle sur des nouvelles données. Le choix de répartition se fait d'une manière aléatoire, ce choix peut présenter des limites. En effet, les données sélectionnées sur la base d'apprentissage peuvent être majoritairement des bons crédits (ou l'inverse), et c'est pareil pour les variables explicatives.

Pour cela, on utilisera la Cross Validation. Cette méthode est équivalente à une répétition de la méthode précédente. En effet, on divise la base en  $n$  sous-groupes, à chaque fois on construit le modèle sur  $n-1$  groupe, puis on le teste sur  $n$ -ième. Par conséquent on aura  $n$  modèles.

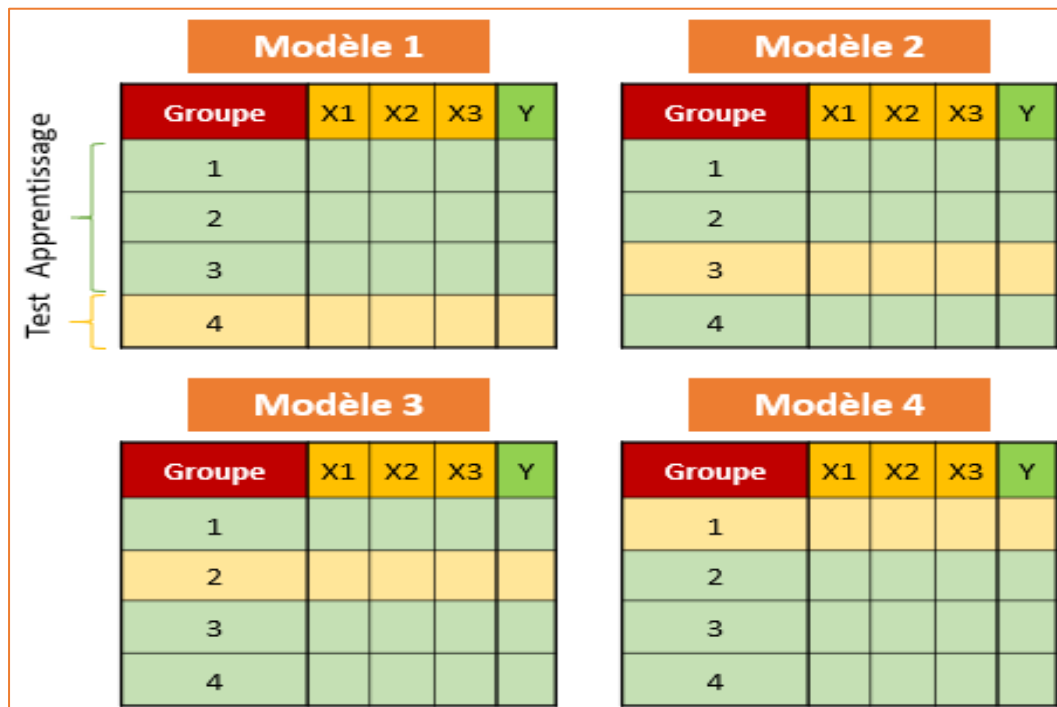


Figure 23 : Illustration de la technique de Cross Validation

### 5.3. Matrice de confusion

La matrice de confusion est une table de contingence entre les valeurs réelles et les valeurs prédites par un modèle. Cette matrice permet de mesurer la qualité d'un modèle de classification binaire :

		Y réelle	
		1	0
Y prédite	1	VP	FN
	0	FP	VN

Tableau 2 : Matrice de confusion

Vrais Positifs (VP) : les 1 prédits correctement

Vrais Négatifs (VN) : les 0 prédits correctement

Faux Négatifs (FN) : les 0 prédits incorrectement

Faux Positifs (FP) : les 1 prédits incorrectement

A partir de ces éléments, on calcule un taux global de la qualité de prédiction :

$$\text{taux de bonne prédictions} = \frac{VP+VN}{\text{somme}}$$

On calcule également calculer le taux de 1 correctement prédits. Ce taux s'appelle sensibilité et

$$\text{on a : } \text{sensitivité} = \frac{VP}{VP+FP}$$

Et finalement on développe le taux des 0 prédits correctement, ce taux s'appelle spécificité, et

$$\text{on a : } \text{Spécificité} = \frac{VN}{VP+FP}$$

### 5.4. Courbe de ROC et AUC

La courbe ROC est la courbe représentant la sensibilité en fonction de 1-la spécificité en faisant évaluer le seuil de probabilité à partir duquel on considère qu'une observation peut porter le label positif. Le choix des seuils n'est pas aléatoire, il utilise les probabilités prédites et les différentes valeurs quelle prennent en les utilisant tour à tour comme seuil.

L'AUC représente l'aire sous la courbe de ROC, plus l'AUC est proche de 1, plus le modèle testé est performant.

# Partie III : Application & Modélisation

## 1. Traitement de la base de données

La première étape de l'analyse consiste à préparer une base contenant la variable défaut ainsi que des variables explicatives nécessaires.

La base de données reçue est constituée de plusieurs bases mensuelles à partir de janvier 2017 jusqu'à mars 2020. Ces bases indiquent la situation des crédits à la fin de chaque mois. Elles sont accumulatives, c'est-à-dire que la base du mois (M) contient en plus des crédits octroyés durant ce mois, tous les crédits enregistrés dans les mois précédents, même s'ils sont soldés.

Chaque base contient 80 variables : des variables d'identification de l'agence (Code agence, code gestionnaire...), des variables d'identification du client (Code, nom et numéro de téléphone ...), des variables concernant la situation sociale du client (Sexe, Situation matrimoniale, et milieu de résidence ...), des variables renseignant la situation financière du client (Activité, Revenu mensuel, et les soldes de ses comptes bancaires ...), des variables constantes durant la vie du prêt (Le montant décaissés, la durée du prêt, et le montant de l'échéance ...), et finalement, des variable indiquant la situation du prêt à la fin du mois concerné (le statut du prêt, l'encours, le portefeuille à risque, le nombre d'incidents commis dans la période et le nombre de jours d'impayé ...).

A partir ce dernier type de variable, et puisque la variable "défaut" n'est pas déterminée sur ces bases, on calculera le défaut le défaut sur un retard de remboursement d'une mensualité de 30 jours.

En plus du traitement de cette variables, on vérifie la cohérence des autres variables.

L'unité statistique dans la base finale aura les caractéristiques suivantes :

- Le prêt est de type individuel
- La période du prêt soit comprise entre janvier 2017 et mars 2020 : pour avoir une visibilité sur tout l'historique du crédit.
- Puisqu'il s'agit d'un modèle à l'octroi, toutes les variables prises dans la base sont des variables à renseigner avant l'octroi du crédit sauf la variable défaut, qui est estimée en analysant le comportement du client durant toute la durée du prêt.

On constate une certaine hétérogénéité entre les produits dans les prêts individuels, le premier type de produit concerne les faibles montants, par contre les deux autres produits concernent des montants plus élevés. Pour cette raison, on sépare ces deux types et on construit un modèle pour chaque type.

Puisque les deux modèles seront construits de la même façon, on se restreint à la modélisation du premier type seulement dans ce mémoire.

Dès le départ, on élimine certaines variables qui sont mal renseignées ou ne représentent pas de variations telles que :

Variable	Description	Problèmes
<b>Ville</b>	La ville de résidence du client	Cette variable n'est pas renseignée pour 43% des prêts
<b>Localisation</b>	Le milieu de résidence du client : urbain ou rural	Seulement 13 sur 4448 observations ont la modalité rurale
<b>Type de logement</b>	Propriétaire, locataire ou autres	88% des observations sont de type "Autres"
<b>Revenu mensuel</b>	Le revenu mensuel du client	98% des observations dont le revenu est nul
<b>Activité du client</b>	L'activité du client	Plus de 90 modalités différentes Difficulté de segmenter ces modalités
<b>Objet de financement</b>	Objet de financement	Plus de 95% des observations ont la même modalité "Trésorerie"
<b>Taux d'intérêt</b>	Le taux annuel du crédit	Existence de 2 taux, l'un des deux est considéré dans 85% des crédits
<b>Sexe</b>	Le sexe du client	Certaines réglementations, ainsi que les valeurs de notre client, interdisent de tarifier ses produits ou de noter les clients selon des variables du genre, de la race et de la religion ..., pour cela cette variable ne serait pas incluse dans la modélisation.

Tableau 3 : variables utiles pour l'analyse, mais mal renseignées sur la base de données

Trois variables sont extraites de la base sans changements ou traitement, il s'agit de :

Variable	Description
<b>Age</b>	L'âge du client au moment de la demande du crédit
<b>Echéance</b>	Montant de la mensualité à payer. Cette variable est fortement corrélée avec plusieurs variables telles que le montant du prêt, les frais d'assurance et les frais de dossier ...
<b>Durée</b>	La durée du prêt en jours

Tableau 4 : Variables tirées de la base sans traitements

Et finalement, quelques autres variables nécessitaient des traitements avant de les exploiter :

Variable	Description	Traitement
<b>Défaut</b>	$\begin{cases} 1 & \text{en cas de Défaute} \\ 0 & \text{sinon} \end{cases}$	<p>Cette variable est calculée pour un retard de paiement d'une mensualité de 30 jours.</p> <p>On observe l'historique du prêt sur toute sa durée, et on enregistre le défaut si on détecte un remboursement qui dépasse 30 jours.</p>
<b>Sit.Mat</b>	La situation matrimoniale du client (Célibataire, marié, divorcé ou veuf)	Les effectifs de divorcés et des veufs sont très faibles, pour cela on les regroupe avec les mariés.
<b>Zone</b>	La zone géographique de l'agence où le client est enregistré	<p>Les variables concernant le milieu de la résidence du client telles que la ville et la localisation ... ne sont pas bien renseignées, par contre, le code d'agence est bien contrôlé et obligatoire dans la saisie des données.</p> <p>On suppose que cette zone peut être considérée comme zone de résidence sous l'hypothèse que les clients sont enregistrés dans les agences les plus proches.</p> <p>On segmente les agences en 4 classes selon le degré du risque.</p>
<b>Nb.Cycles</b>	Nombre de cycles : Le nombre de crédits octroyés au client précédemment	<p>Segmentation des clients selon de classes :</p> <ul style="list-style-type: none"> <li>• Premier crédit : les clients qui demandent un crédit pour la première fois</li> <li>• Plusieurs : les clients dont la banque a déjà octroyé un crédit ou plus</li> </ul>
<b>Solde 1</b>	La moyenne des soldes des fins des mois du compte à vue du client chez la banque	<p>Le manque de variables qui renseignent sur la situation (financière et budgétaire) du client telles que le revenu et les dépenses, nous pousse à analyser ses flux à travers ses comptes chez la banque</p> <p>Le solde dans un moment donné, ne reflète pas la situation réelle, pour cela on calcule une moyenne</p>

			des soldes sur une période donnée.
<b>Cap.Remb.1</b>	Capacité de remboursement du Solde1	de	Le ratio du Solde1 sur le montant de l'échéance L'utilisation de ce ratio nous permettrait de comparer la capacité de remboursement des clients, ainsi que leurs situations. Par exemple, un client dont la Cap.Remb.1 est égale à 1,5, cela veut dire qu'en moyen, ce client il aura 150% du montant de l'échéance dans son compte à la fin du mois. Ce client serait moins risqué qu'un client avec une Cap.Remb.1 égale à 0,8, même si ce dernier a le plus grand solde1
<b>CV.S1</b>	Coefficient de variation du Solde1		Ecart-type du Solde1 sur sa moyenne. Cette variable sert à analyser la stabilité du solde1
<b>Solde 2</b>	La moyenne des soldes des fins des mois du compte d'épargne du client chez la banque		Même traitement du Solde1 Le Solde2 concerne un compte d'épargne, la banque a le droit d'extraire le montant de l'échéance à partir de ce compte si le Solde1 n'est pas suffisant.
<b>Cap.Remb.2</b>	Capacité de remboursement du Solde2	de	Même traitement de la Cap.remb.1
<b>Caution</b>	Le ratio du Solde Caution sur le montant de l'échéance		Le solde Caution est un montant payé par le client comme garantie La variable Caution est calculée pour avoir une échelle comparative entre les clients
<b>CV.S2</b>	Coefficient de variation du Solde2		Même traitement de la CV.S1

Tableau 5 : Variables traitées

Cette partie nous confirme les difficultés relatives à la microfinance. Elles concernent la qualité des données, ainsi que la quantité des observations et des variables importantes.

## 2. Analyse exploratoire des variables

### 2.1. Variables quantitatives

La base finale contient 4 variables qualitatives y compris le défaut, leurs distributions est comme suit :

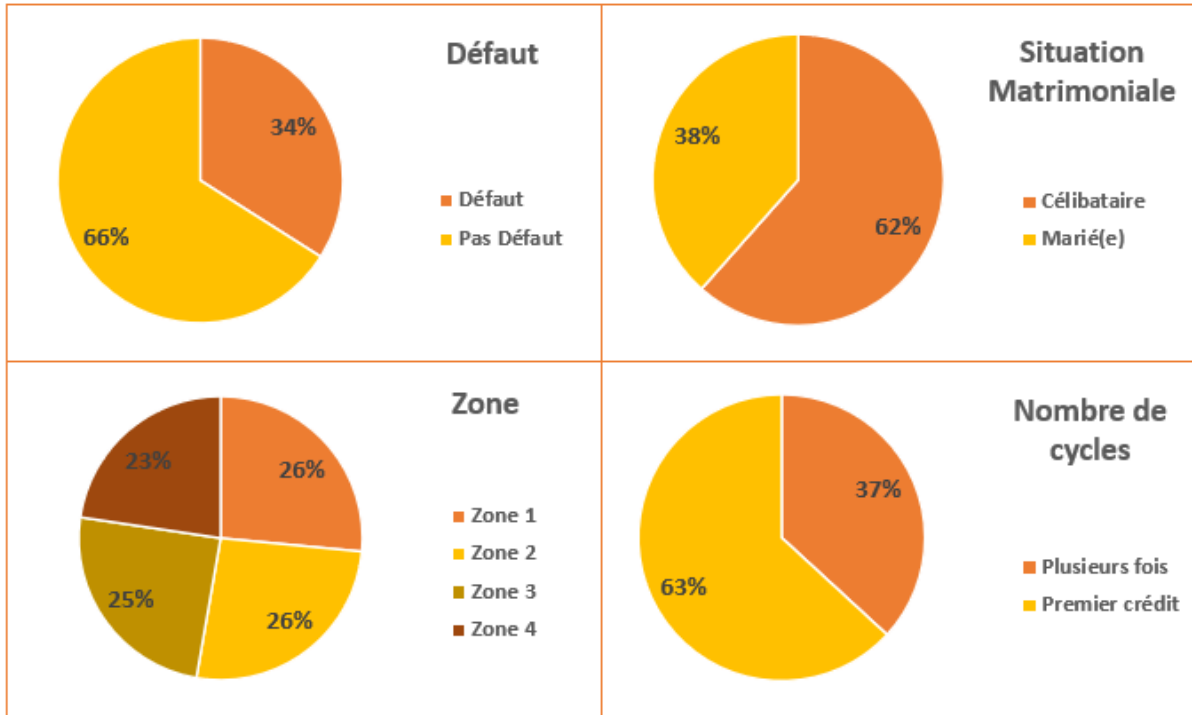


Figure 24 : Distribution des variables qualitatives

D'après la figure, le taux de défaut est de 34% pour ce produit, le portefeuille contient plus de crédits de la première fois et les individus sont distribués sur les zones d'une façon identique.

Ensuite, on construit des tables de contingence entre ces variables pour mesurer les effets discriminatifs entre elles :

Variable	Taux de défaut	Test Khi-2	Conclusion
<b>Sit.Mat</b>	Célibataire	36%	$\chi^2=9,59$ DF=1 P-value =0,0019
	Mariré	31%	
<b>Zone</b>	Zone1	19%	$\chi^2=230,38$ DF=3 P-value < 2,2.e-16
	Zone2	30%	
	Zone3	38%	

	Zone4	52%		seuil 5%
<b>Nb.Cycles</b>	Premier crédit	30%	$\chi^2=42,74$ DF=1 P-value =6,25.e-11	On rejette l'hypothèse de l'indépendance entre la variable Nb.Cycles et Défaut au seuil 5%
	Plusieurs fois	41%		

Tableau 6 : résumé des tables de contingence des variables qualitatives

D'après l'analyse des tables de contingences de ces variables avec la variable "Défaut", on constate :

- Les clients célibataires ont tendance à commettre plus de défauts que les mariés.
- La variable Zone sépare les clients selon des classes de risque où le taux de défaut est considérablement différent entre ces classes.
- Les clients qui demandent un 2<sup>ième</sup> crédit ou plus sont plus risqués que ceux qui se présentent pour la première fois. Ce résultat apparaît contradictoire avec la politique de la banque à l'octroi de nouveaux crédits pour les anciens clients, où elle accepte les bons profils en se basant sur la performance dans les anciens prêts. Ce résultat peut être interprété par le changement du comportement de ces clients puisqu'ils sont plus habitués aux processus et connaissent les différentes pénalités ...

Pour tester statistiquement les effets de ces variables, on procède par un test de Kh-2. D'après les résultats du tableau, les trois variables ont un effet discriminatif sur la variable "Défaut" au seuil de 5%.

## 2.2. Variables qualitatives

### 2.2.1. Corrélations

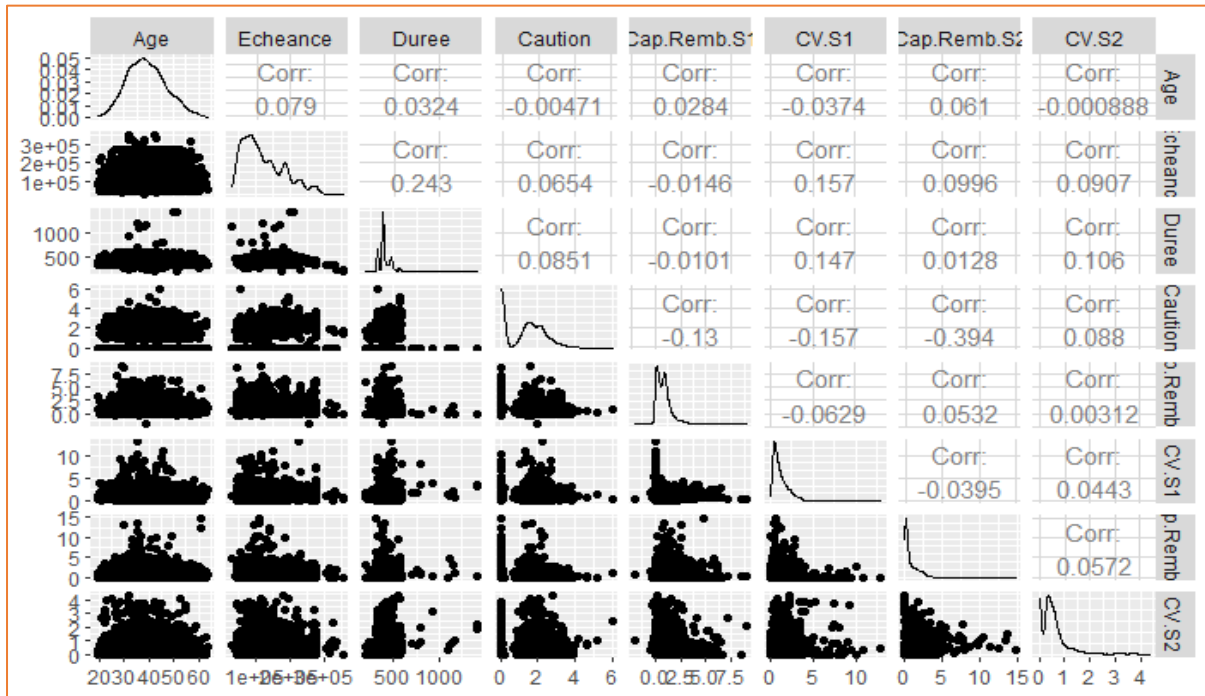


Figure 25 : matrice des corrélations des variables quantitatives

D'après la figure, la plus grande corrélation se trouve entre les variables "Echéance" et "Durée" avec un coefficient de 24%, ce qui permet de confirmer que les variables du modèle sont décorrélées entre elles.

Ce constat est confirmé par les nuages des points entre les variables, où ils ne suivent aucune tendance pour toutes les combinaisons.

### 2.2.2. Effets sur le Défaut

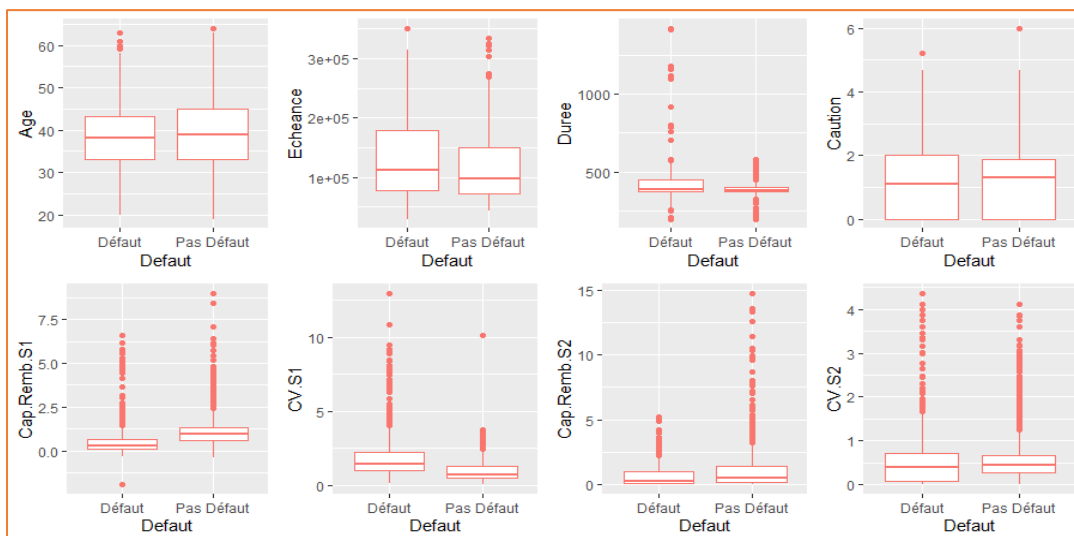


Figure 26 : Boîtes à moustache des variables quantitatives en fonction du défaut

L'analyse graphique de ces boîtes à moustache nous permet de pronostiquer les effets des variables quantitatives sur le défaut :

- Age : les individus moins âgés ont tendance à commettre plus de défauts que les individus les plus âgés
- Echéance : les prêts dont les montants sont plus grands tombent en défaillance plus que les prêts à faibles montants
- Durée : pareille à la variable "Echéance", les prêts étalés sur des longues durées tombent en défaillance plus que les prêts étalés sur des courtes durées
- Caution : les clients qui ne paient pas de caution, ont tendance à commettre plus de défauts
- Cap.Remb.1 : la différence pour cette variable est très claire graphiquement, les clients dont la capacité de remboursement sur le compte à vue est proche de zéro ont tendance à commettre plus de défaut. C'est-à-dire que ces individus n'ont pas en le montant de la mensualité dans leurs comptes.
- CV.S1 : Pour cette variable, la différence est également claire. En effet, les clients avec un solde instable ou volatile (coefficient de variation élevé) ont tendance à commettre plus de défaut.
- Cap.Remb.2 : de même que pour la variable "Cap.Remb.1", les clients dont la capacité de remboursement sur le compte d'épargne est proche de zéro ont tendance à commettre plus de défaut.
- CV.S2 : de même que pour la variable "CV.S1", les clients avec un solde instable ou volatile (coefficient de variation élevé) ont tendance à commettre plus de défaut.

Pour tester ces constats statistiquement, on procède par test de Student pour la comparaison des moyennes sur les deux groupes :

Variable	Moyenne Totale	Moyenne "Défaut"	Moyenne "Pas Défaut"	Test de Student	Conclusion
Age	39,05	38,74	39,34	t=-3,1569 DF=2543,5 P-value=0,0016	On rejette l'hypothèse d'égalité des moyennes des deux groupes au niveau de 5%

<b>Echéance</b>	120088	126696,8	116684,4	t=4,7145 DF=2426,4 P-value=2,56.e-06	On rejette l'hypothèse d'égalité des moyennes des deux groupes au niveau de 5%
<b>Durée</b>	395,4	409,76	388,03	t=7,8519 DF=1645,4 P-value=7,33.e-15	On rejette l'hypothèse d'égalité des moyennes des deux groupes au niveau de 5%
<b>Caution</b>	1,107	1,047	1,137	t=-2,3789 DF=2300,9 P-value=0,0174	On rejette l'hypothèse d'égalité des moyennes des deux groupes au niveau de 5%
<b>Cap.Remb.1</b>	0,891	0,523	1,080	t=-18,576 DF=2385,6 P-value < 2,2.e-16	On rejette l'hypothèse d'égalité des moyennes des deux groupes au niveau de 5%
<b>CV.S1</b>	1,269	1,818	0,986	t=20,747 DF=1598,6 P-value < 2,2.e-16	On rejette l'hypothèse d'égalité des moyennes des deux groupes au niveau de 5%
<b>Cap.Remb.2</b>	0,863	0,647	0,974	t=-9,173 DF=3379,8 P-value < 2,2.e-16	On rejette l'hypothèse d'égalité des moyennes des

					deux groupes au niveau de 5%
<b>CV.S2</b>	0,552	0,542	0,5572	t=-0,706 DF=2095,1 P-value=0,4798	On accepte l'hypothèse d'égalité des moyennes des deux groupes au niveau de 5%

Tableau 7 ; tests de Student d'égalité des moyennes des variables qualitatives des deux populations : défauts et sains

D'après ce tableau, on peut avoir d'idées sur les variables qui ont un effet sur le défaut :

- Variables à grand effet : Cap.Remb.1, CV.S1 et Cap.Remb.2.
- Variables à moyen effet : Age, Echéance, Durée.
- Variable à faible effet : Caution
- Absence d'effet : CV.S2

### Conclusion

En guise de conclusion, l'analyse des liaisons entre les variables nous permet de sélectionner les variables nécessaires pour la modélisation ; elles sont décorrélées entre elles, et qui ont un effet discriminatif sur la variable "Défaut".

Ces variables sont : Zone, Cap.Remb.1, CV.S1, Cap.Remb.2, Nb.Cycles, Age, Echéance, Durée, Caution et Sit.Mat.

### 3. Modélisation

Afin de tester la performance des modèles, on sépare la base de données aléatoirement en deux parties :

- Base d'apprentissage : cette base contient 75% (effectif de 2651) des observations sur lesquelles les modèles seront construits
- Base de test : cette base contient les 25% (effectif de 891) restantes des observations de la base complète, cette base servira à tester et valoriser la performance des modèles sur des nouvelles données.

Le taux de défaut est le même sur les deux bases : 34%

#### 3.1. Régression logistique

##### 3.1.1. Premier modèle

A l'aide de la fonction "glm" sous R, on construit un premier modèle d'apprentissage sur les variables choisies précédemment

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.693e-01  6.625e-02  2.555  0.0107 *
Age          -2.461e-03  1.031e-03  -2.386  0.0171 *
Sit.MatMarié(e) -1.448e-02  1.701e-02  -0.852  0.3945
ZoneZone 2    5.476e-02  2.270e-02  2.412  0.0159 *
ZoneZone 3    1.411e-01  2.253e-02  6.261  4.44e-10 ***
ZoneZone 4    2.319e-01  2.384e-02  9.727  < 2e-16 ***
Nb.CyclesPremier crédit -3.354e-02  1.690e-02  -1.984  0.0473 *
Echeance      5.250e-08  1.389e-07  0.378  0.7054
Duree         5.882e-04  1.225e-04  4.803  1.65e-06 ***
Caution      -3.729e-02  8.754e-03  -4.260  2.12e-05 ***
Cap.Remb.S1  -1.331e-01  9.218e-03 -14.435  < 2e-16 ***
CV.S1         1.340e-01  7.845e-03  17.080  < 2e-16 ***
Cap.Remb.S2  -6.765e-02  7.639e-03  -8.856  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.1635625)

Null deviance: 596.68  on 2650  degrees of freedom
Residual deviance: 431.48  on 2638  degrees of freedom
AIC: 2738.4

```

Figure 27 : Premier modèle de RL

L'analyse des déviations nous permet de conclure que le modèle est adéquat. On constate également que les paramètres des deux variables "échéance" et "Sit.Mat=Marié" sont significativement nuls au seuil de 5%.

Ensuite on utilise la méthode "Stepwise" pour sélectionner seulement les variables qui apportent plus d'informations sur le modèle.

### 3.1.2. Modèle optimal

La sélection des variables à l'aide de la méthode "Stepwise" en minimisant le critère AIC nous a permis d'exclure les variables "échéance" et "Sit.Mat" du modèle.

La régression logistique sur les variables restantes nous a donné les résultats suivants :

Coefficients:				
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.1741156	0.0660329	2.637	0.00842 **
Age	-0.0026298	0.0010031	-2.622	0.00880 **
ZoneZone 2	0.0514602	0.0223819	2.299	0.02157 *
ZoneZone 3	0.1410064	0.0225109	6.264	4.37e-10 ***
ZoneZone 4	0.2333450	0.0237763	9.814	< 2e-16 ***
Nb.CyclesPremier crédit	-0.0345762	0.0168288	-2.055	0.04002 *
Duree	0.0005943	0.0001202	4.943	8.17e-07 ***
Caution	-0.0367678	0.0086983	-4.227	2.45e-05 ***
Cap.Remb.S1	-0.1331604	0.0092150	-14.450	< 2e-16 ***
CV.S1	0.1344188	0.0077670	17.306	< 2e-16 ***
Cap.Remb.S2	-0.0674377	0.0075699	-8.909	< 2e-16 ***

---  
 Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
 (Dispersion parameter for gaussian family taken to be 0.1634914)

Null deviance: 596.68 on 2650 degrees of freedom  
 Residual deviance: 431.62 on 2640 degrees of freedom  
 AIC: 2735.2

Figure 28 : Modèle optimale de la RL

On constate que tous les paramètres du modèle sont significativement non nuls au seuil de 5%

Afin de mesurer la performance prédictive de ce modèle sur la base d'apprentissage, on dessine la courbe de ROC, puis on calcule la surface sous la courbe.

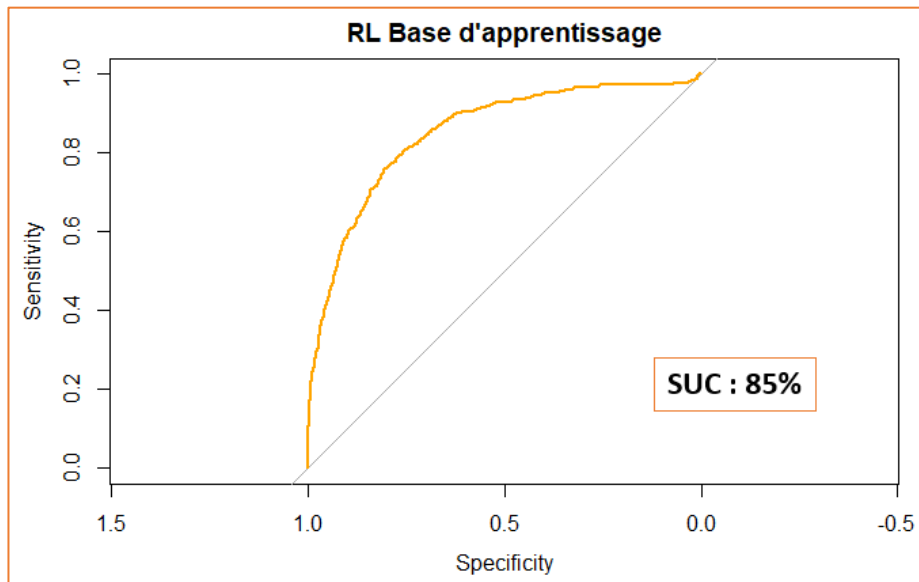


Figure 29 : Courbe de ROC de la RL sur la base d'apprentissage

On a eu une SUC de 85%, Par conséquent on peut considérer que la qualité de d'ajustement du modèle est acceptable.

### 3.1.3. Les effets des variables sur la probabilité de défaut

Afin de comparer les effets des variables sur la probabilité de défaut, on construit des modèles en régressant la variable "Défaut" sur chaque variable individuellement. Ensuite, on trace les courbes de ROC pour chaque modèle.

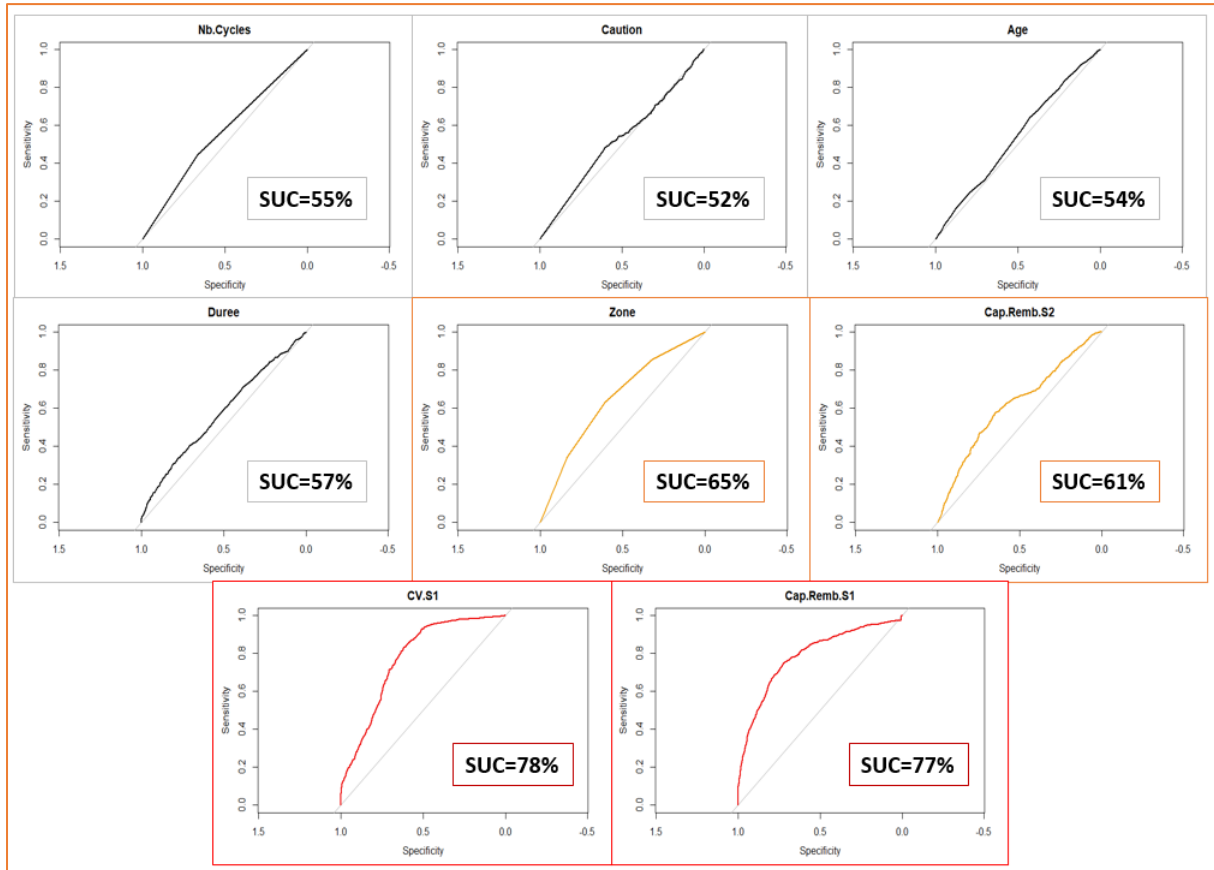


Figure 30 : Courbes de ROC des RL des variables du modèle su le défaut

La comparaison des surface sus la courbe de ROC entre les modèles, nous permet de classer les variables selon leurs effets sur le défaut :

- Variables à effet élevé : Cap.Remb.1 et CV.S1
- Variables à effet moyen : Cap.Remb.2 et Zone
- Variables à effet faible : Age, Durée, Caution et Nb.Cycles

Ces résultats étaient prévisibles dès l'analyse exploratoire des variables où on a trouvé presque les mêmes classes.

Pour estimer la tendance de la probabilité de défaut en variant les variables explicatives, on procède par analyse des signes des paramètres dans le modèle obtenu :

- Variable quantitative : un signe négatif (resp. positif) signifie que la probabilité de défaut est une fonction décroissante (resp. croissante) par rapport à cette variable.

- Variable qualitative : un signe négatif (resp. positif) signifie que la probabilité de défaut est plus faible (resp. élevée) dans la modalité en question que la modalité de référence.

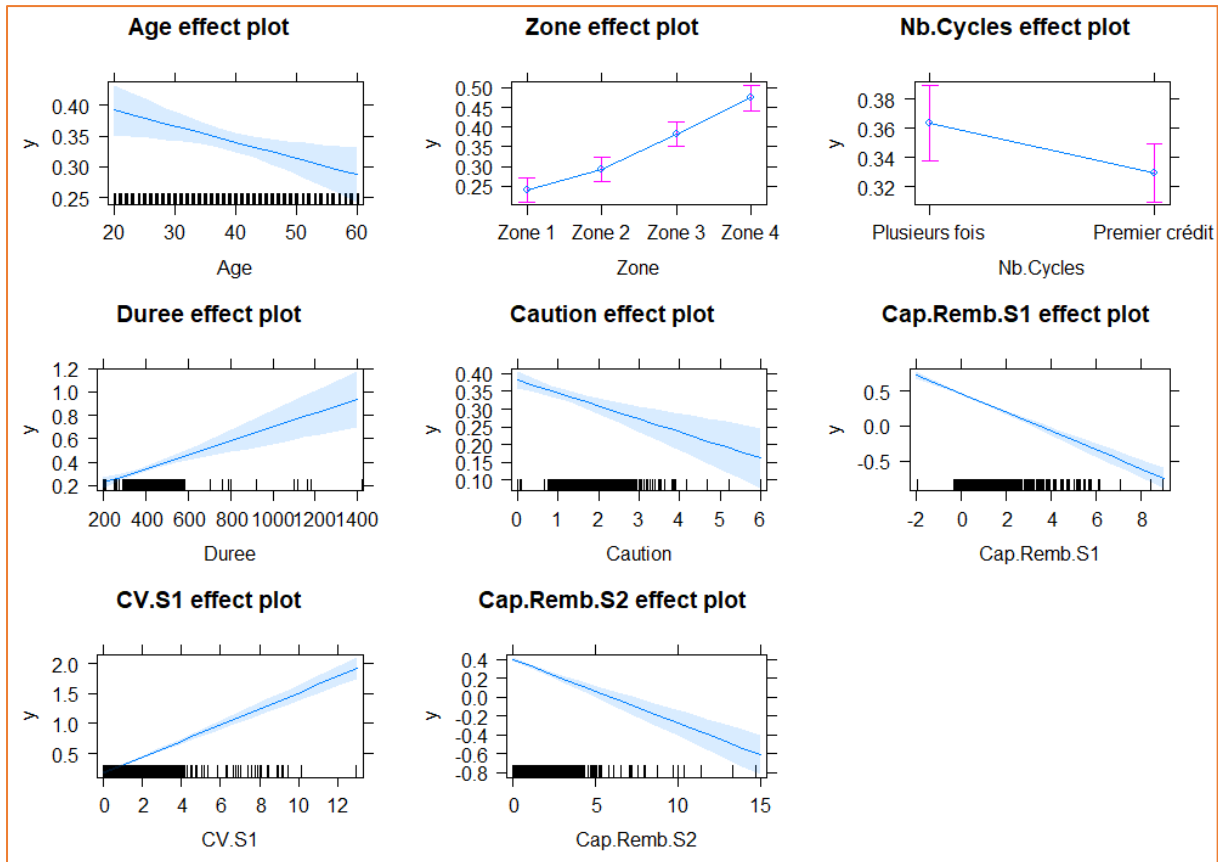


Figure 31 : Courbes des effets des variables sur le défaut

La figure résume le sens de variation du  $\log(\text{Odds})$  de la probabilité de défaut, et par conséquent de la probabilité de défaut, et on a :

- La probabilité est une fonction décroissante de
  - Age : le risque de défaut est plus faible chez les clients les plus âgés
  - Caution : le paiement d'une Caution élevée par rapport au montant de l'échéance diminue la probabilité de défaut
  - Cap.Remb.1 et Cap.Remb.2 : les clients qui ont des capacités de remboursement plus élevées ont une grande probabilité à honorer leurs engagements
- La probabilité est une fonction croissante de
  - Durée : les crédits étalés sur de longues durées ont une probabilité de défaut plus grande.
  - CV.S1 : les clients dont le Solde.1 varie significativement, c'est-à-dire qu'il est instable risquent de ne pas rembourser leurs mensualités.

- La probabilité de défaut chez les clients qui demandent un crédit pour la première fois est plus faibles.

Ces résultats sont également cohérents avec l'analyse exploratoire des variables.

### 3.1.4. Détermination du seuil de prédiction

En général, on utilise un seuil de probabilité de 0,5 pour déterminer les valeurs prédites par le modèle. En d'autres termes, si la probabilité estimée est supérieure à 0,5, on classifie l'observation comme défaut, et dans le cas contraire, on la classifie comme saine.

La figure suivante représente les probabilités de défaut estimées par le modèle logistique appliqué sur la base d'apprentissage.

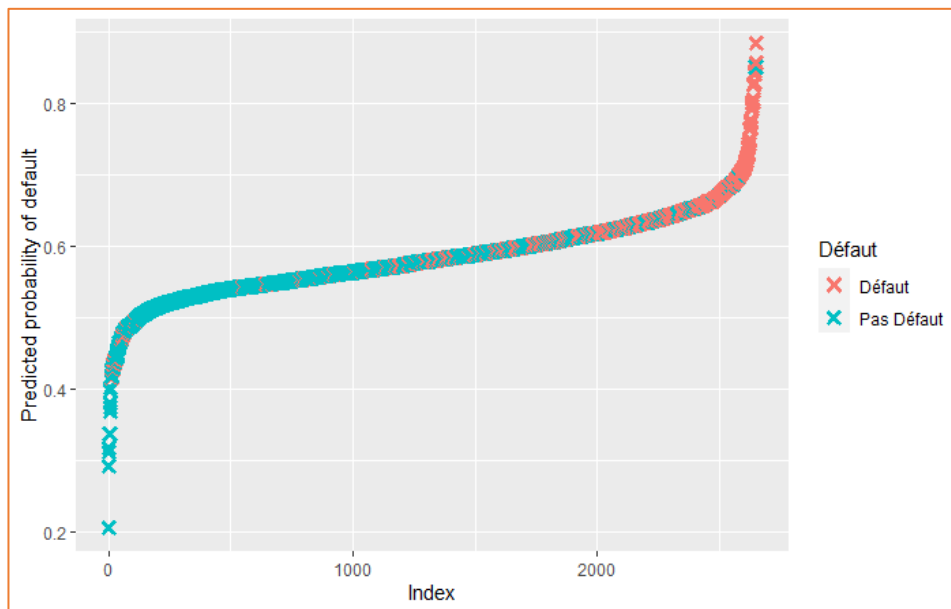


Figure 32 : probabilités de défaut estimées par la RL

Graphiquement, le seuil qui permettra une prédiction ne correspond pas à 0.5, le seuil optimal se trouve dans les environs de 0.6 où la distinction entre les crédits en défaut et les crédits sains est plus claire.

Pour chercher le meilleur seuil, on construit des matrices de confusion en fonction de la probabilité qui permettront de calculer les taux de performance de la prédiction du modèle :

- Erreur de prédiction : le taux des observations classifiées incorrectement (1-accuracy)
- Taux des Faux Positifs : le taux des crédits sains classifiés en défaut (1-Sensitivité)
- Taux des Faux Négatifs : le taux des crédits en défaut classifiés comme sains (1-Spécificité)

Ensuite on trace ces taux en fonction des probabilités :

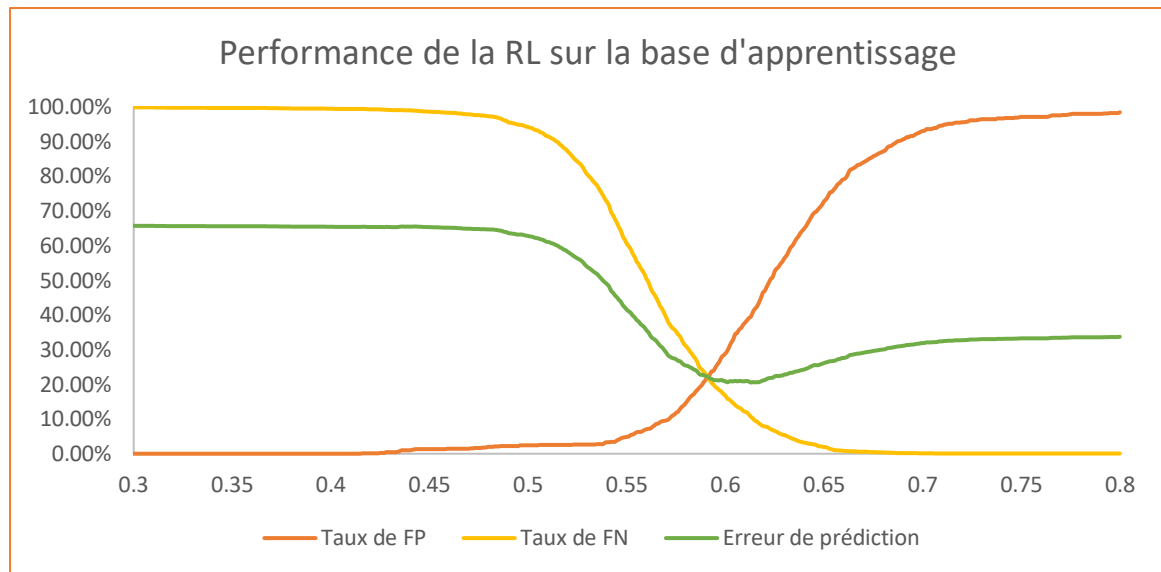


Figure 33 : les erreurs de performance de la RL sur la base d'apprentissage

Idéalement, on a intérêt à minimiser les trois taux simultanément. Pour l'instant, on s'intéresse par le seuil qui minimise l'erreur de la prédiction.

Numériquement, ce seuil correspond à la probabilité 0,613. La matrice de confusion correspondante est comme suit

		Référence	
		Défaut	Pas Défaut
Prédiction	Défaut	548	185
	Pas Défaut	359	1559

Tableau 8 : matrice de confusion de la RL sur la base d'apprentissage pour une probabilité de 0,5

D'après cette matrice, les taux sont :

- L'erreur de prédiction est égale à 20% (Accuracy = 80%)
- Le taux des faux positifs est égal à 40% (Sensitivité = 60%)
- Le taux des faux négatifs est égal à 10% (Spécificité = 90%)

Le taux des FP est élevé pour ce seuil. En effet, le but principal de la modélisation, consiste à prévoir les clients qui vont tomber en défaillance de remboursement. Avec ce seuil, 40% des individus qui ont commis réellement un défaut ne sont pas détectés par le modèle. Il est préférable de construire des classes de risque qui contiennent des degrés du risque de défaut et non pas seulement la prédiction de deux classes (Défaut, Pas défaut).

Le point d'interaction des trois courbes permet d'avoir le même niveau d'erreur sur les trois. Ce point correspond à la probabilité de 0,59 qui donne une erreur de 22%. Donc ce seuil améliore considérablement le taux des FP tout en gardant la même grandeur du taux d'Accuracy (78%).

## 3.2. Arbre de classification

### 3.2.1. Construction de l'arbre

Le deuxième modèle utilisé pour la prédiction de la probabilité de défaut est l'arbre de classification CART.

La modélisation sur la base d'apprentissage nous a permis d'avoir l'arbre suivant :

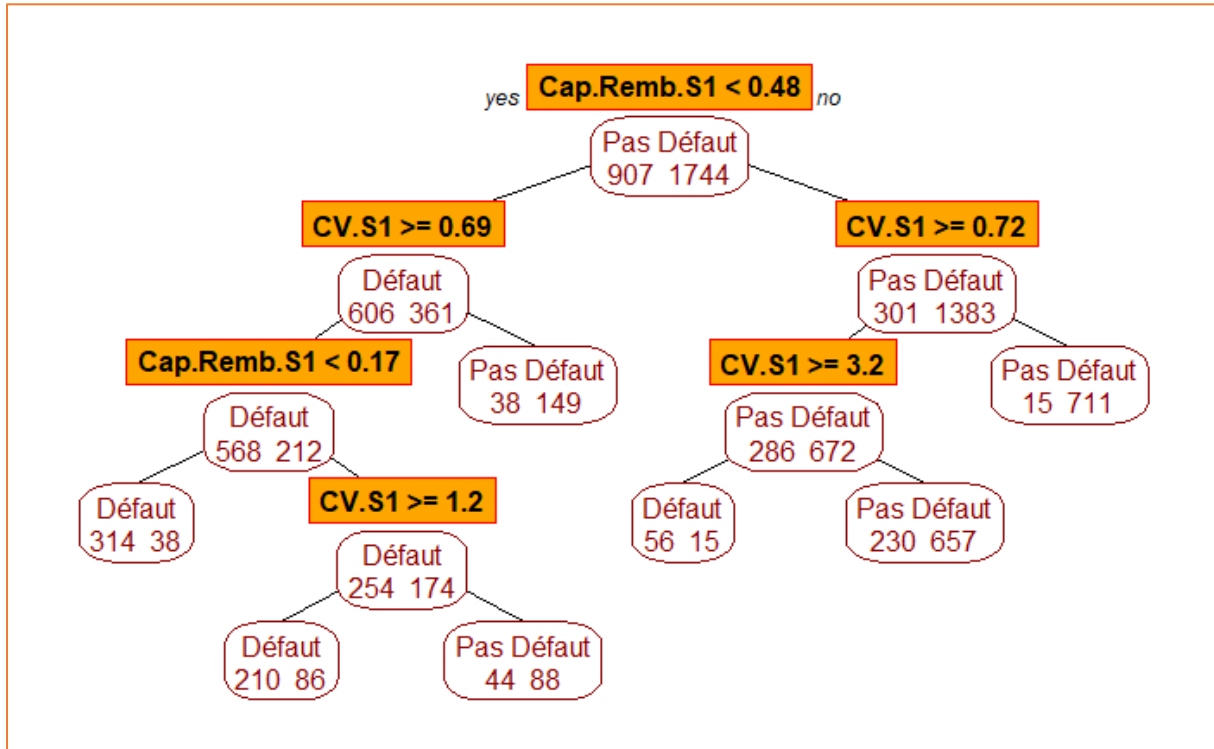


Figure 34 : Arbre de décision obtenue par le modèle Arbre de classification

L'arbre est constitué de 6 nœuds intermédiaires et 7 feuilles ou régions terminales. Les variables qui participent à la construction de l'arbre sont : Cap.Remb.S1 et CV.S1. D'après le modèle de la régression logistique, ces deux variables avaient le plus grand effet sur la probabilité de défaut.

### 3.2.2. Estimation des probabilités de défaut

A la base, le modèle de l'arbre de classification n'estime pas les probabilités de défaut, il construit directement des régions de prédiction.

Mais, on peut estimer des probabilités de défaut comme ratio des prêts en défaut sur l'effectif total dans chaque feuille.

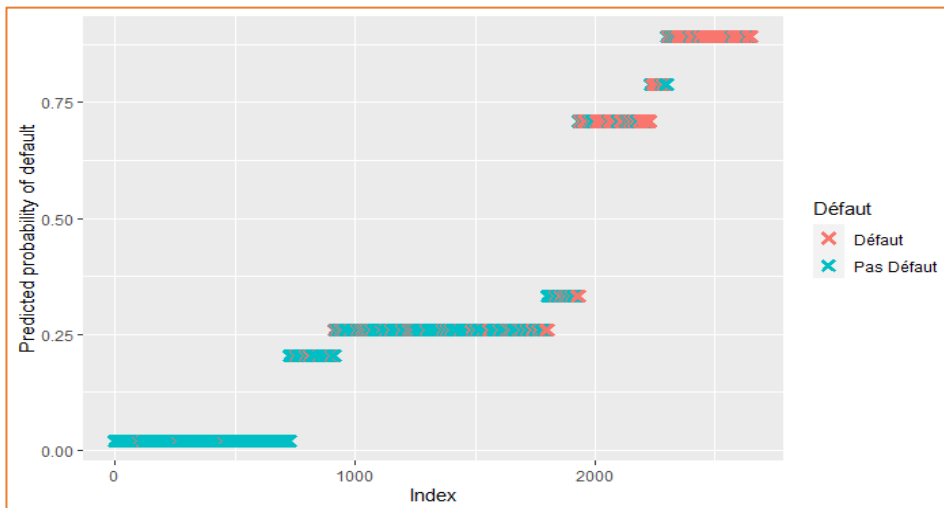


Figure 35 : Probabilités de défaut estimées pour l'arbre de classification

Il est clair que le seuil utilisé pour la prédiction est 0,5. Mais on peut également construire des classes de risque pour ce type de modèles à l'aide de ces probabilités.

### 3.2.3. Analyse de la performance sur la base d'apprentissage

A travers ces probabilités, on peut construire la courbe de ROC pour évaluer la performance du modèle.

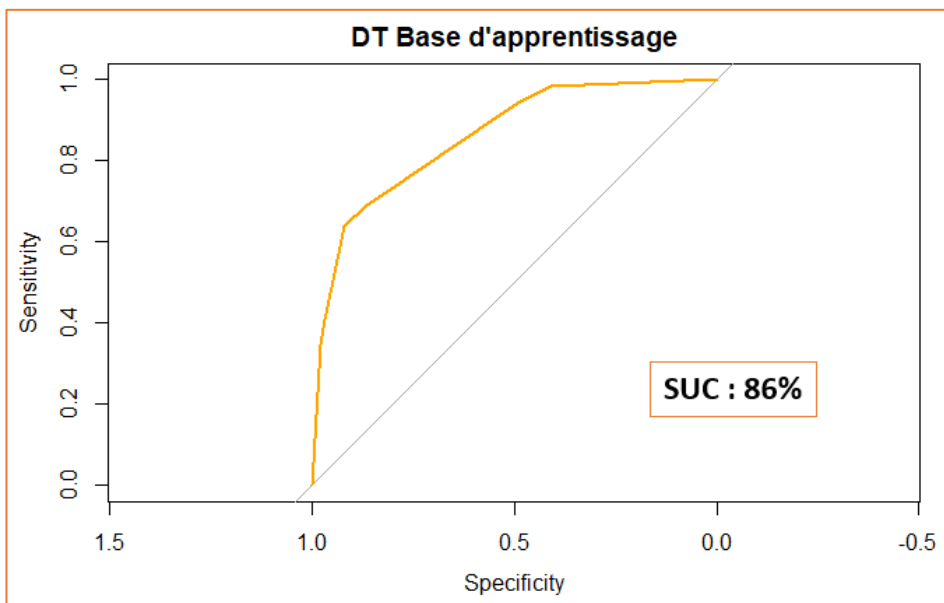


Figure 36 : Courbe de ROC de l'arbre de classification sur la base d'apprentissage

La courbe de ROC construite à partir de la base d'apprentissage nous donne une surface sous la courbe égale à 86%.

La matrice de confusion correspondante est :

		Référence	
		Défaut	Pas Défaut
Prédiction	Défaut	580	139
	Pas Défaut	327	1605

Tableau 9 : Matrice de Confusion de l'arbre de classification sur la base d'apprentissage

D'après cette matrice, les taux sont :

- L'erreur de prédiction est égale à 18% (Accuracy = 82%)
- Le taux des faux positifs est égal à 36% (Sensitivité = 64%)
- Le taux des faux négatifs est égal à 8% (Spécificité = 92%)

### 3.3. Forêt aléatoire

#### 3.3.1. Choix du nombre d'arbre dans le modèle

Initialement, on construit une forêt avec 500 arbres. Le choix du nombre suffisant d'arbre consiste à minimiser les erreurs de prédiction : l'erreur de l'OOB (Accuracy), le taux des FP (de défauts classifiés incorrectement) et le taux des FN (Taux de sains classifiés incorrectement).

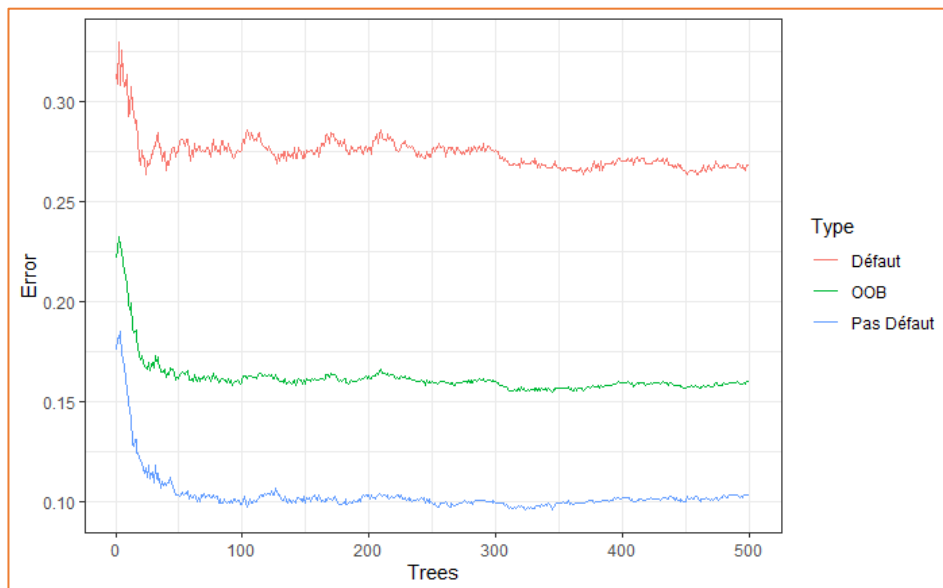


Figure 37 : les taux d'erreurs commises par le RF en fonction du nombre d'arbres

La figure nous montre qu'à partir d'un nombre d'arbres de 300, ces trois taux se stabilisent et ne peuvent plus diminuer. Donc une forêt de 500 arbres est largement suffisante

#### 3.3.2. Choix du nombre de variables à considérer dans chaque répartition

La construction des arbres par Random Forest se restreint sur un nombre réduit de

variables tirées aléatoirement pour la constitution des tests dans les nœuds.

Pour choisir le meilleur nombre de variable à tirer parmi les 10 variables choisies pour la modélisation, on construit des forêts aléatoires en changeant ce paramètre de 1 à 10.

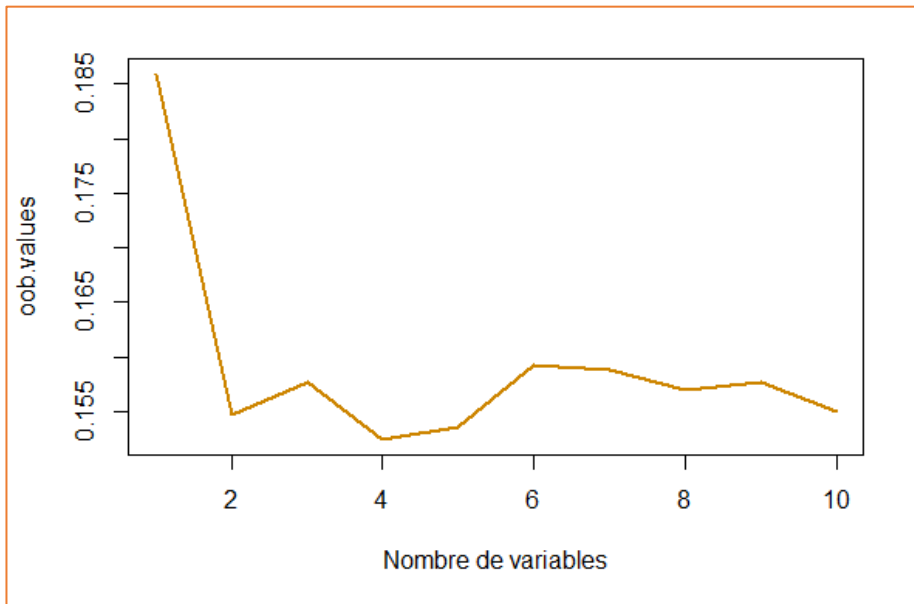


Figure 38 : Variation de l'erreur d'OBB en fonction du nombre de variables considérées dans chaque répartition

D'après la figure à côté, la sélection de 4 variables permet de minimiser l'erreur de l'OOB.

Dans la suite de l'analyse, on crée un Random Forest avec un nombre d'arbres égal à 500 et un nombre de 4 variables tirées dans chaque étape.

### 3.3.3. Effets des variables

Pour évaluer les effets des variables sur le modèle, on mesure la capacité totale de diminution de l'impureté des nœuds par la répartition d'une variable.

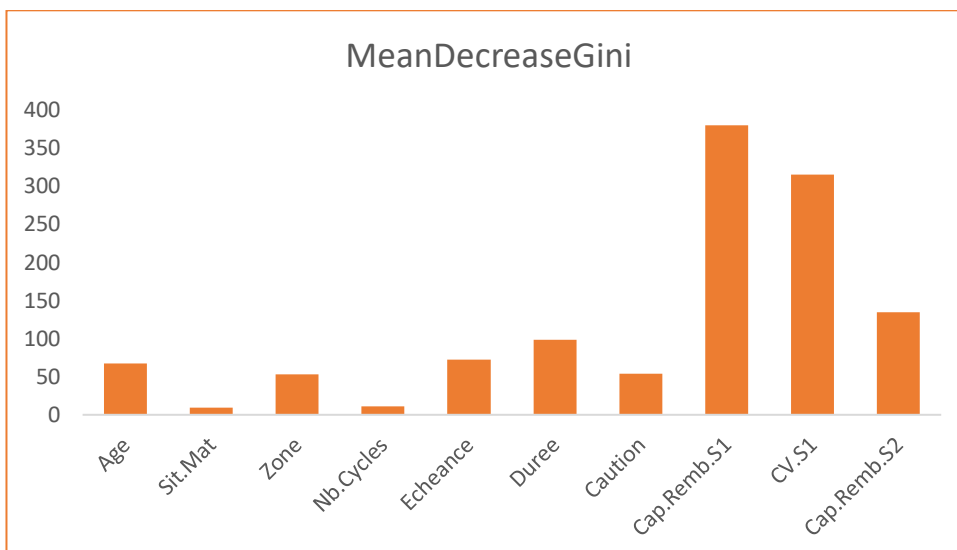


Figure 39 : Effets des variables sur le RF

D'après la figure, les variables qui ont les plus grandes effets sur le défaut sont Cap.Remb.S1 et CV.S1, ce qui rejoint les conclusions des autres modèles et de l'analyse exploratoire.

### 3.3.4. Probabilités de défaut estimées

A partir des votes des arbres pour la prédiction des classes, on peut construire une probabilité de défaut en divisant le nombre d'arbres qui ont voté "Défaut" sur le nombre total des arbres (500 dans ce modèle).

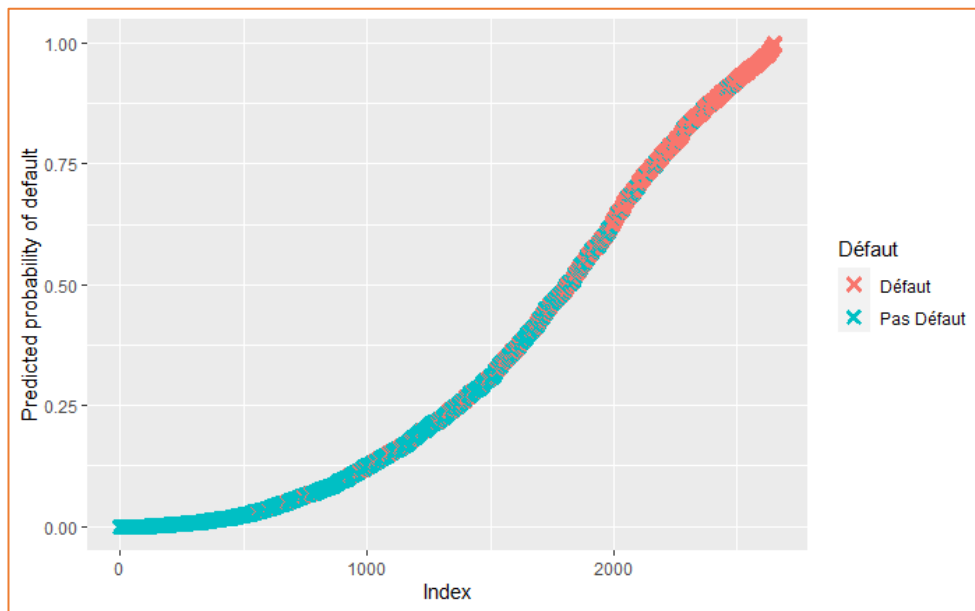


Figure 40 : Probabilités de défauts estimées sur la base d'apprentissage pour le modèle RF

### 3.3.5. Performance du modèle sur la base d'apprentissage

A travers ces probabilités estimées, on peut tracer la courbe de ROC du modèle et calculer la surface sous cette courbe :

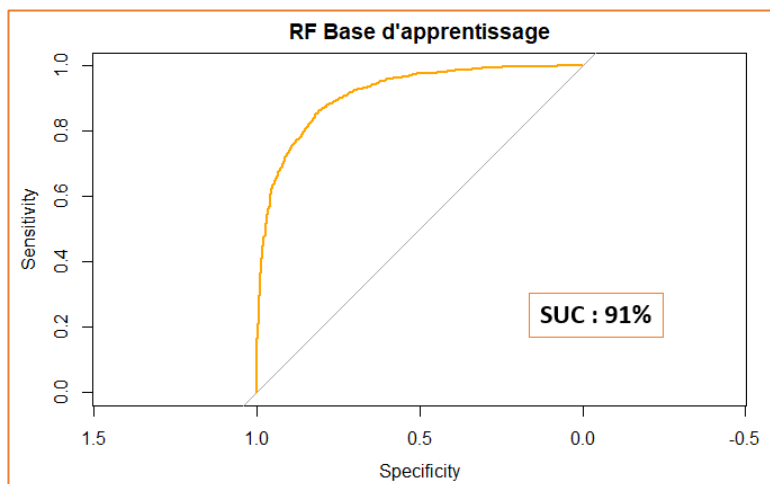


Figure 41: Courbe de ROC du modèle RF sur la base d'apprentissage

La surface sous la courbe de ROC est égale à 91%

Ensuite on construit la matrice de confusion sur cette base :

		Référence	
		Défaut	Pas Défaut
Prédiction	Défaut	668	170
	Pas Défaut	239	1574

Tableau 10 : Matrice de confusion du modèle RF sur la base d'apprentissage

D'après cette matrice, les taux sont :

- L'erreur de prédiction est égale à 15% (Accuracy = 85%)
- Le taux des faux positifs est égal à 26% (Sensitivité = 74%)
- Le taux des faux négatifs est égal à 10% (Spécificité = 90%)

### 3.4. Gradient Boosting

#### 3.4.1. Choix des paramètres du modèle

Pour construire un modèle GBM, il faut déterminer quelques inputs tels que le nombre d'arbres et le taux d'apprentissage (Learning Rate).

Les valeurs par défaut de ces deux paramètres dans la fonction *gbm* sous R sont respectivement 100 et 0.1

Un taux d'apprentissage plus faible augmente la qualité des prédictions, en contrepartie, il nécessite plus d'arbres.

On choisit un taux d'apprentissage de 0.01, cela nécessite un grand nombre d'arbres. Initialement on construit un modèle avec 10000 arbres :

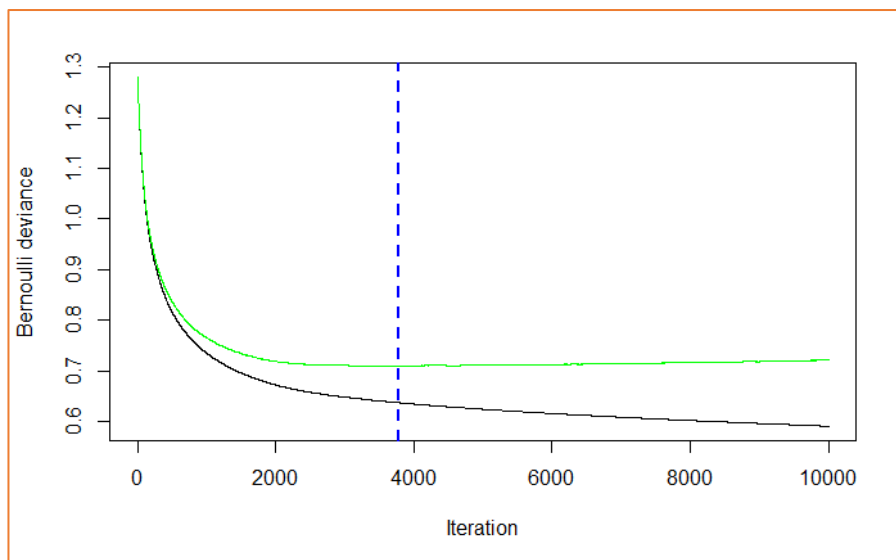


Figure 42 : sélection du nombre d'arbre optimal pour un taux d'apprentissage de 0.01 par la méthode de Cross Validation

La figure à côté nous indique que le nombre optimale d'arbres à utiliser est d'environ 4000 et plus précisément dans cet exemple 3779 (la ligne bleue).

Ce nombre est offert par une comparaison entre des erreurs d'apprentissage (courbe noire) et les erreurs de test par 10 classes de Cross-validation ( Courbe verte). À partir de se seuil, on risque d'avoir un phénomène de surajustement.

À titre d'exemple, on construit un modèle avec un taux d'apprentissage de 0.1 avec 1000 arbre.

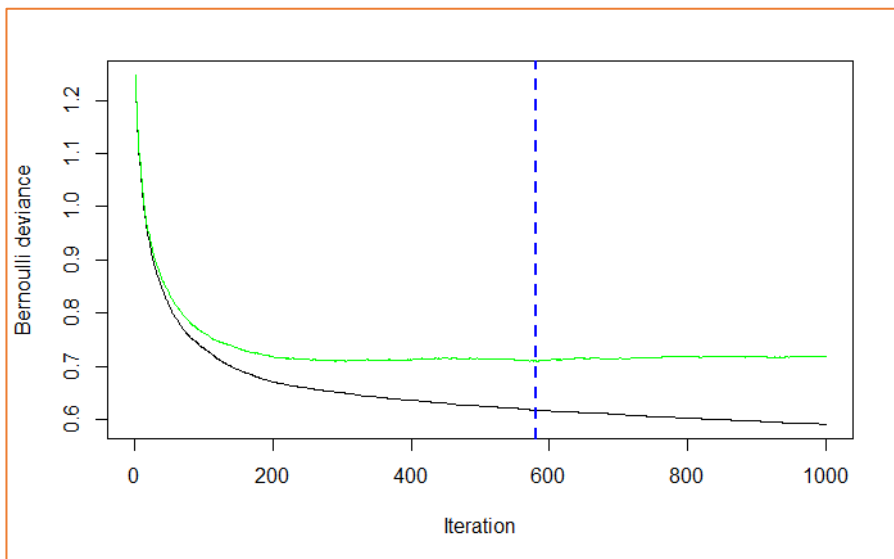


Figure 43 : sélection du nombre d'arbre optimal pour un taux d'apprentissage de 0.1 par la méthode de Cross Validation

D'après la figure à côté, le nombre optimale d'itérations est d'environ 600. Cela confirme la nécessité d'un plus grand nombre d'arbre en cas de diminution du taux d'apprentissage.

### 3.4.2. Les effets des variables sur le modèle

Pour mesurer l'importance des variables dans la construction du modèle, on calcule le nombre d'utilisation de chaque variable dans les séparations des nœuds.

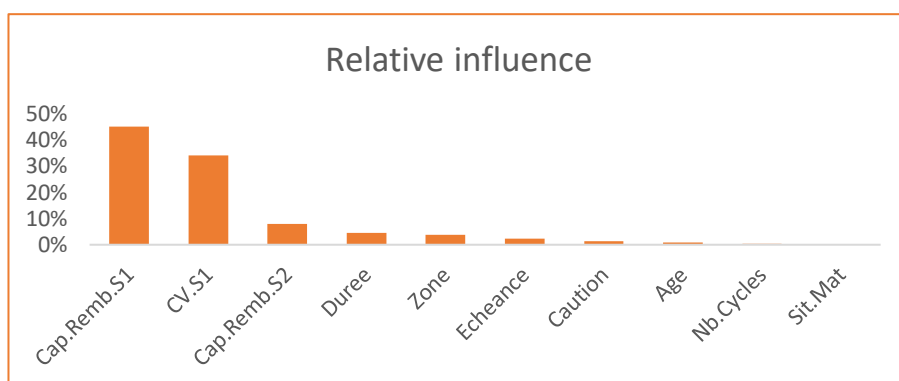


Figure 44 : Relative Influence des variables sur le GBM

D'après le graphe, la variable Cap.Remb.S1 est utilisée dans les 45% des répartitions, ensuite la variable est utilisée dans 34% des séparations.

Le Gradient boosting confirme également que ces deux variables sont les plus importantes pour l'estimation de la probabilité de défaut.

### 3.4.3. Probabilité de défaut

Comme dans le modèle de la régression logistique, les modèles du Gradient Boosting estiment des probabilités de défaut à partir desquelles on prévoit le défaut lorsqu'elles dépassent un certain seuil.

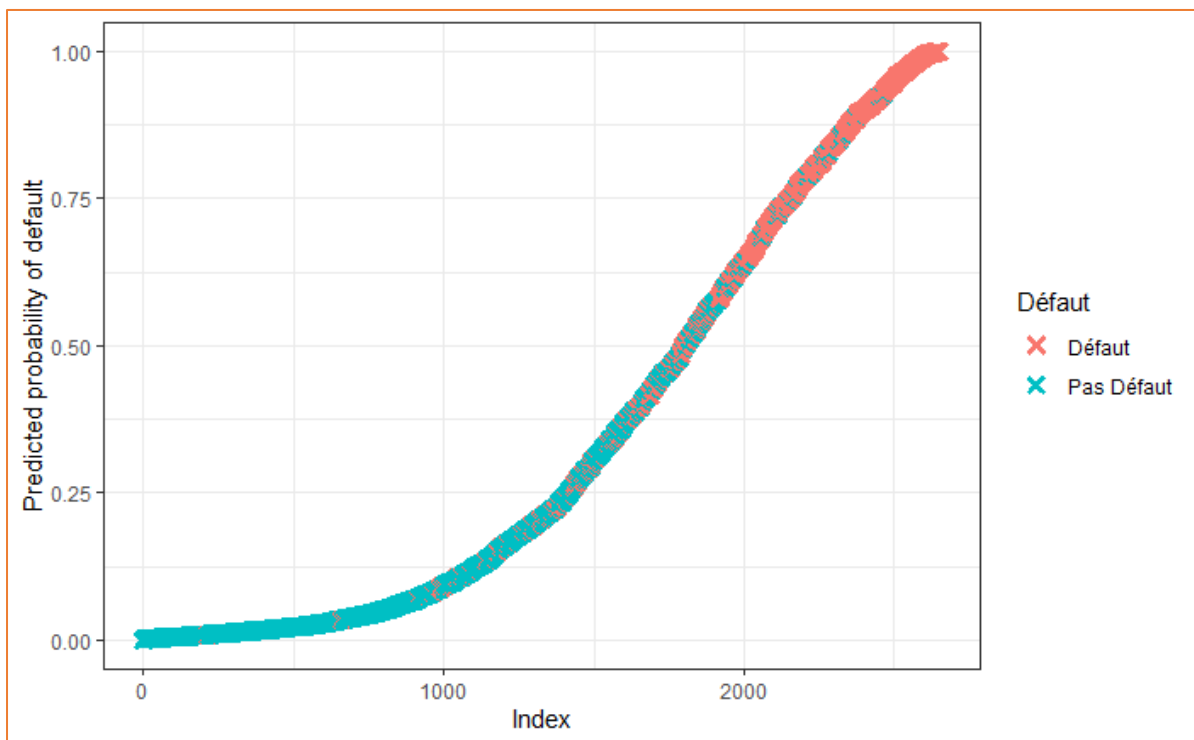


Figure 45 : Probabilités de défaut prédites pour le modèle GBM sur la base d'apprentissage

Graphiquement, le seuil optimal peut varier entre 0.3 et 0.7 en fonction de la stratégie de la banque, si la banque est très averse au risque, elle peut choisir un seuil faible pour détecter le maximum des défauts, en contrepartie elle perdra des bons profils classés incorrectement comme défaut.

### 3.4.4. Performance du modèle sur la base d'apprentissage

La courbe de ROC pour ce modèle sur la base d'apprentissage nous donne une surface sous la courbe égale à 93%

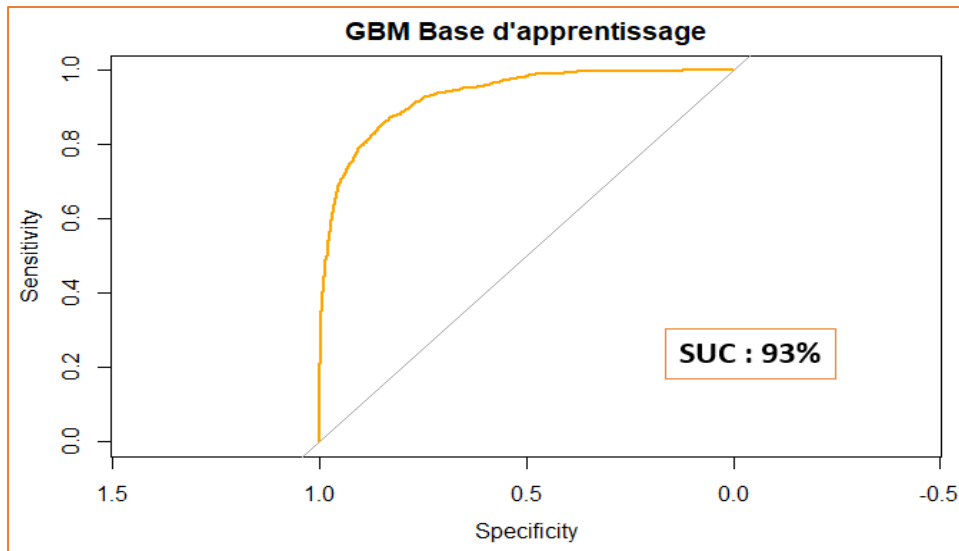


Figure 46 : Courbe de ROC pour le modèle GBM sur la base d'apprentissage

Ensuite, on trace les trois taux d'erreurs en fonction des seuils de prédictions :

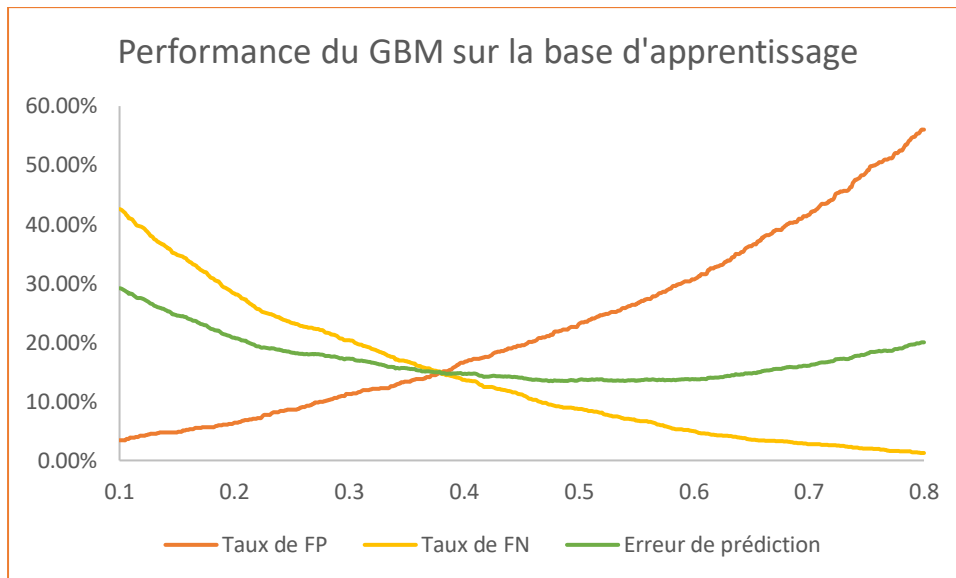


Figure 47 : Performance du modèle GBM sur la base d'apprentissage

Comme dans le cas de la régression logistique, on distingue entre 2 seuils :

- Le seuil d'interaction des trois courbes qui correspond à 0.38 avec une erreur de 14.75%
- Le seuil qui minimise l'erreur de prédiction, il correspond à 0.48, sa matrice de confusion est :

		Référence	
		Défaut	Pas Défaut
Prédiction	Défaut	709	198
	Pas Défaut	160	1584

Tableau 11 : Matrice de confusion du modèle GBM sur la base d'apprentissage pour une probabilité de 0.48

D'après cette matrice, les taux sont :

- L'erreur de prédiction est égale à 14% (Accuracy = 86%)
- Le taux des faux positifs est égal à 22% (Sensitivité = 78%)
- Le taux des faux négatifs est égal à 9% (Spécificité = 91%)

## 4. Comparaisons des modèles

### 4.1. Base d'apprentissage

Avant de passer à la validation sur la du test, on rappelle les performances des modèles sur la base d'apprentissage :

Modèle	Probabilité	Bonnes prédictions	Sensitivité	Spécificité
Régression logistique	P=0.613	79.48%	60.42%	89.39%
	P=0.591	77.78%	77.78%	77.78%
Arbre de classification		82.42%	63.95%	92.03%
RF		84.57%	73.65%	90.25%
GBM	P=0.48	86.50%	78.17%	90.83%
	P=0.38	85.25%	85.25%	85.25%

Tableau 12 : Indicateurs de performance des 4 modèles sur la base d'apprentissage

Le tableau montre que le modèle est le plus performant, surtout pour un seuil de prédiction de 0.38 où les trois erreurs sont de l'ordre de 15%. Pour le deuxième meilleur modèle, on hésite entre le RF et la RL avec un seuil de 0.591. Cela dépend de l'aversion au risque du décideur, s'il veut détecter le maximum de défaut, alors il doit choisir la RL, s'il veut maximiser l'accuracy, le choix est le RF.

### 4.2. Validation sur la base test

Initialement, on compare les courbes de ROC des quatre modèles : le meilleur modèle est le Random Forest avec une surface sous la courbe est presque 93%, suivi par le modèle dy Gradient Boosting avec une surface de 92%.

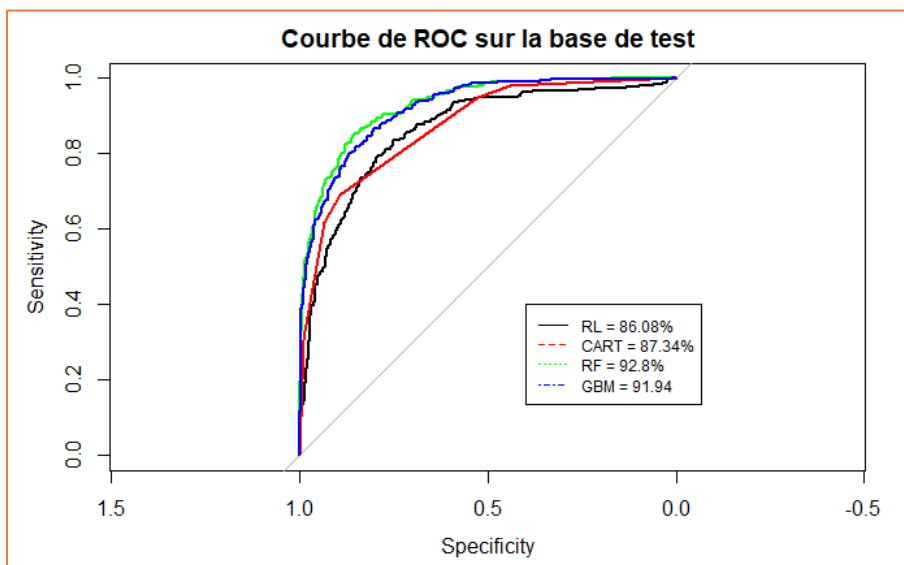


Figure 48 : Courbes de ROC des 4 modèles sur la base de test

Ensuite, on calcule les taux de performance :

Modèle	Probabilité	Bonnes prédictions	Sensitivité	Spécificité
Régression logistique	P=0.613	79.04%	60.03%	89.24%
	0.591	79.01%	77.78%	79.63%
Arbre de classification		82.83%	61.95%	93.27%
RF		86.2%	71.72%	93.43%
GBM	P=0.38	84.40%	74.07%	89.56%
	P=0.38	83.95%	80.81%	85.52%

Tableau 13 : Indicateurs de performance des 4 modèles sur la base de test

Le RF est le meilleur modèle concernant le taux de bonnes prédictions ;

Le modèle GBM avec seuil de 0.38 assurent une meilleure combinaison des 3 indicateurs.

### 4.3. Cross Validation

Dans cette partie, on procède par une Cross Validation sur 10 classes : on crée 10 groupes à partir de la base complète, à chaque fois on construit le modèle sur 9 groupes et on le teste sur le dixième. Finalement on construit un modèle moyen entre les moyens obtenus.

#### 4.3.1. Régression logistique

Initialement, sur les 10 modèles construits, la méthode de sélection des variables Stepwise a éliminé les deux variables "échéance" et "Sit.Mat.".

	Est. 1	Est. 2	Est. 3	Est. 4	Est. 5	Est. 6	Est. 7	Est. 8	Est. 9	Est. 10	Moy
(Intercept)	0,129	0,113	0,110	0,137	0,144	0,133	0,164	0,110	0,132	0,170	0,134
Age	-0,002	-0,002	-0,002	-0,002	-0,002	-0,002	-0,003	-0,003	-0,002	-0,002	-0,002
ZoneZone 2	0,071	0,065	0,065	0,051	0,063	0,063	0,063	0,074	0,049	0,056	0,062
ZoneZone 3	0,138	0,130	0,139	0,131	0,133	0,133	0,136	0,141	0,127	0,125	0,133
ZoneZone 4	0,219	0,224	0,225	0,231	0,233	0,224	0,219	0,232	0,213	0,213	0,223
Nb.CyclesPremier	-0,029	-0,036	-0,037	-0,032	-0,043	-0,044	-0,044	-0,028	-0,038	-0,041	-0,037
Duree	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001
Caution	-0,034	-0,034	-0,034	-0,035	-0,036	-0,032	-0,032	-0,032	-0,038	-0,035	-0,034
Cap.Remb.S1	-0,137	-0,134	-0,138	-0,135	-0,131	-0,144	-0,129	-0,135	-0,147	-0,149	-0,138
CV.S1	0,144	0,138	0,142	0,147	0,146	0,141	0,139	0,141	0,142	0,138	0,142
Cap.Remb.S2	-0,063	-0,064	-0,061	-0,060	-0,061	-0,058	-0,063	-0,064	-0,059	-0,060	-0,061

Tableau 14 : Estimations des paramètres de la RL sur les 10 classes de la Cross Validation

Les estimations des paramètres des paramètres sur les 10 modèles ont le même signe, ainsi que le même ordre de grandeur.

La comparaison entre le modèle établi sur la base d'apprentissage et les moyennes des estimations des modèles obtenus par la Cross Validation, montre des différences considérables.

On considère par exemple le profil suivant :

- Age = 35
- Habite dans la zone 4
- Premier crédit
- Une durée de 365 jours
- Sans Caution
- Cap.Remb.S1=0.7
- CV.S1 = 0.5
- Cap.Remb.S2 = 0.8

La probabilité obtenue par le premier modèle serait 0.65, et la probabilité obtenue par le modèle moyen des modèles de la Cross Validation est 0.56. C'est une différence importante en termes de probabilités, surtout parce qu'elles se trouvent dans la région de détermination du seuil qui minimise le taux d'erreur de prédiction. Dans ce cas, on doit déterminer un seuil pour chaque modèle.

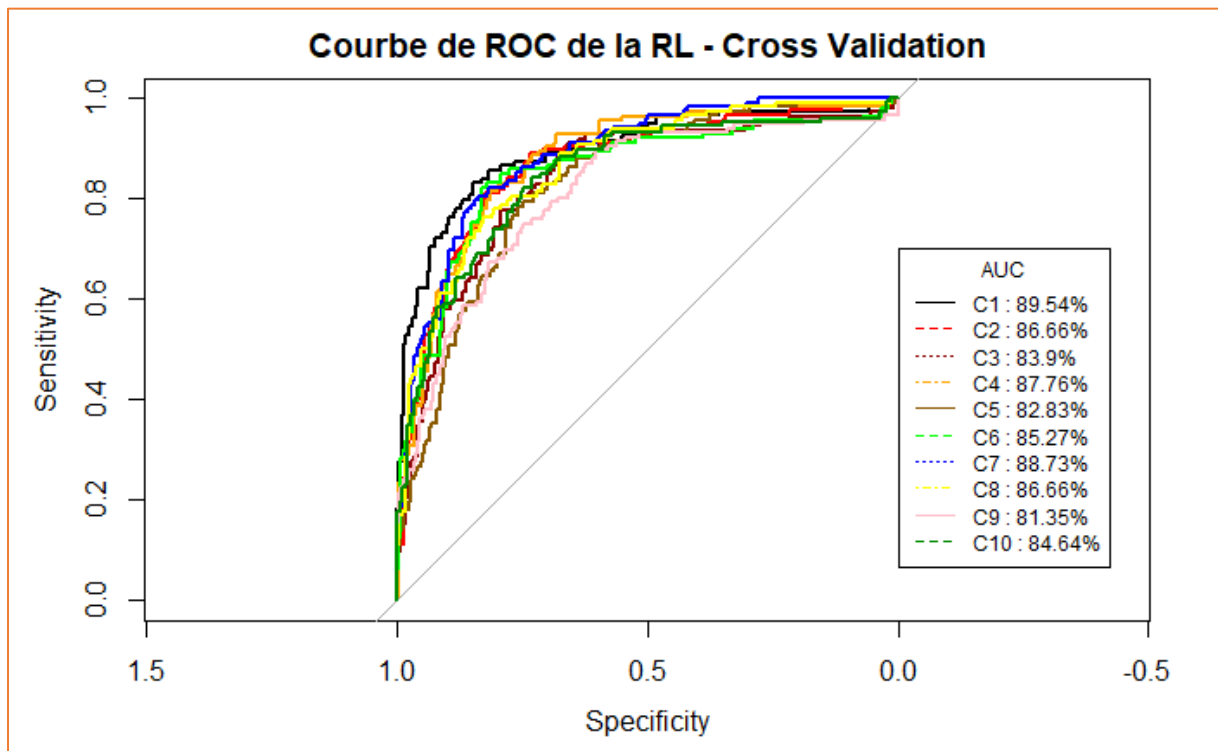


Figure 49 : Courbes de ROC de la RL sur les bases test - Cross Validation 10 classes

La sensibilité par rapport à la base d'entraînement du modèle se voit graphiquement sur les courbes de ROC. Les surfaces sous les courbes varient de 81% jusqu'à 89% avec une surface moyenne de 85.73%.

Ensuite, on mesure la performance de ces modèles par la construction des matrices de

confusion. Ces dernières nous permettent de mesurer le taux de bonnes prédictions, la sensibilité et la spécificité des modèles.

Comme précédemment, on construit ces tables pour deux seuils : le premier minimise l'erreur de prédiction et le deuxième correspond au point d'intersection des trois taux.

Modèle	Probabilité	Bonnes prédictions	Sensitivité	Spécificité
Classe 1	0.612	79.90%	60.71%	89.78%
	0.591	79.11%	78.99%	79.17%
Classe 2	0.612	79.93%	60.63%	89.86%
	0.591	79.05%	78.41%	79.38%
Classe 3	0.612	79.73%	61.54%	89.09%
	0.591	78.88%	79.49%	78.57%
Classe 4	0.612	79.76%	61.88%	89.96%
	0.591	78.09%	78.07%	78.10%
Classe 5	0.612	79.50%	61.54%	88.75%
	0.591	78.15%	78.74%	77.84%
Classe 6	0.612	79.93%	62.13%	89.09%
	0.591	78.63%	79.32%	78.27%
Classe 7	0.612	79.95%	60.88%	89.78%
	0.591	78.97%	78.80%	79.04%
Classe 8	0.612	79.87%	61.05%	89.56%
	0.591	78.57%	78.90%	78.40%
Classe 9	0.612	79.87%	62.21%	89.96%
	0.591	78.46%	79.32%	78.02%
Classe 10	0.612	80.07%	60.88%	89.95%
	0.591	79.05%	78.74%	79.21%
Moyenne	0.612	79.95%	61.38%	89.52%
	0.591	78.97%	79.15%	78.87%

Tableau 15 : Indicateurs de la performance de la RL - Cross Validation 10 Classes

La première remarque concerne les seuils choisis, on a une égalité pour les deux seuils. Ces deux seuils choisis vérifient les conditions discutées sur le modèle moyen.

On constate que seuils presque vérifient ces conditions sur les 10 modèles, pour le premier, les taux de bonnes prédictions sont très proches aux taux de bonnes prédictions des 10 modèles. Pour le deuxième, la différence est négligeable entre les 3 taux pour tous les modèles.

### 4.3.2. Arbre de décision

Sur les 10 arbres<sup>13</sup> construits sont corrélés entre elles, où on constate :

- Les variables utilisées pour les répartitions sont les mêmes : Cap.Remb.S1 et CV.S1.
- L'ordre des séparations par variables est le même sur les 10 arbres (Racine : Cap.Remb.S1, nœuds du premier niveau CV.S1 ...)
- Les seuils de séparation sont presque les mêmes :
  - Racine : le seuil 0.51 est choisi sur 7 arbres et le seuil 0.48 sur le reste avec une moyenne de 0.5
  - Premier nœud à gauche : le seuil 0.79 est choisi sur 5 arbres et le seuil 0.69 est choisi sur 3. La moyenne des seuils pour ce nœud est 0.73.
  - Premier nœud à droite : le seuil 0.69 est choisi sur 4 arbres et le seuil 0.72 est choisi pour 3. La moyenne des seuils pour ce nœud est 0.71.

D'après cet exemple, on constate clairement l'utilité du choix d'un nombre réduit de variable à tenir dans les séparations par les Random Forest. Cette technique permet de construire des arbres décorrélés, ainsi de donner aux autres variables plus de chance pour être sélectionnées dans les séparations.

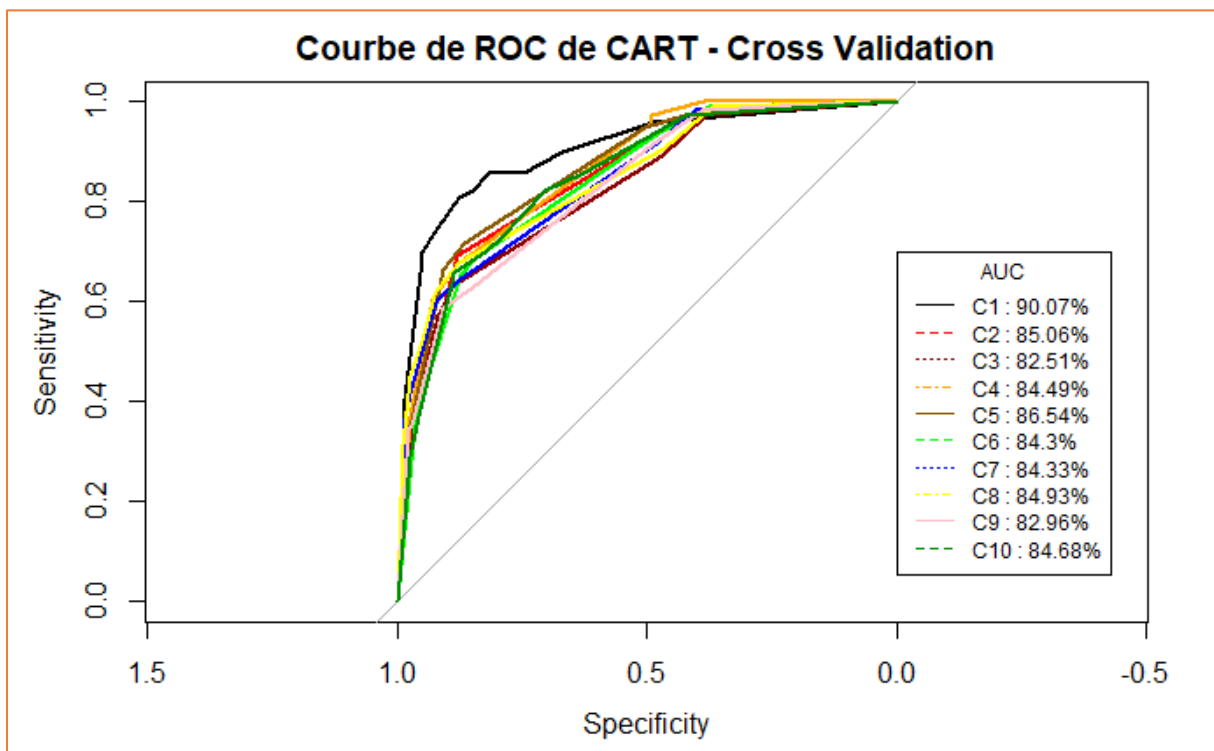


Figure 50 : Courbes de ROC de l'arbre de classification CART sur les bases test - Cross Validation 10 classes

Graphiquement, on constate que la sensibilité par rapport à la base d'entraînement a diminué. A l'exception de la courbe dont la surface est égale à 90.07%, les autres surfaces

<sup>13</sup> Voir Annexe Cross Validation – Arbre de classification

varient entre 82% et 85%.

Contrairement à la régression logistique, l'arbre de classification segmente les observations en premier lieu, puis, on peut calculer les probabilités de défaut. Dans ce type de modèles, on ne rencontre pas le problème de détermination d'un seuil pour la classification, cela veut dire que tous les modèles, construits précédemment ou par la Cross Validation, utilisent les mêmes critères pour la prédiction.

L'arbre de classification diminue la sensibilité par rapport à la base d'entraînement, mais il n'améliore pas beaucoup la qualité d'ajustement, la moyenne des surface sou la courbe de ROC est égale à 85.18% contre 85.73% pour le modèle logistique.

Ensuite, on construit des matrices de confusion pour mesurer la qualité de prédiction du modèle :

Modèle	Bonnes prédictions	Sensitivité	Spécificité
Classe 1	85.8%	73.85%	92.45%
Classe 2	85.99%	70.87%	94.35%
Classe 3	81.9%	70.54%	87.29%
Classe 4	87.17%	74.56%	93.45%
Classe 5	84.82%	73.28%	90.12%
Classe 6	81.79%	69.03%	88.29%
Classe 7	84.2%	74.80%	88.93%
Classe 8	85%	72.88%	90.91%
Classe 9	81.71%	71.55%	86.75%
Classe 10	82.46%	71.92%	88.98%
Moyenne	84.08%	72.32%	90.15%

Tableau 16 : Indicateurs de la performance de l'arbre de classification CART - Cross Validation 10 Classes

D'après le tableau, le taux des bonnes prédictions varie entre 81% et 86% avec une moyenne de 84%, la sensibilité varie entre 69% et 75% avec une moyenne de 72%, et la spécificité varie entre 86% et 94% avec une moyenne de 90%.

En moyen, ces indicateurs sont améliorés par rapport au premier modèle construit, cela est peut-être à cause du changement de l'effectif de la base d'entraînement. Dans le premier, le modèle s'entraîne sur 75% des observations, par contre, les modèles de Cross Validation s'entraînent sur 90% des données.

### 4.3.3. Random Forest

Par construction, le modèle Random Forest est équivalent à une Cross Validation sur un

nombre de groupe égale au nombre d'arbres utilisés. En outre, elle permet une de construire des arbres plus décorrélés entre elles.

Les modèles construits dans cette étape utilisent 1000 arbres. Les graphes<sup>14</sup> d'amélioration des erreurs nous montrent que ce nombre est largement suffisant.

L'ordre des valeurs importante pour la modélisation est le même sur tous les modèles. La moyenne des indicateurs mesurant cette performance, indique que les variables les plus importantes sont la Cap.Remb.S1 et CV.S1 :

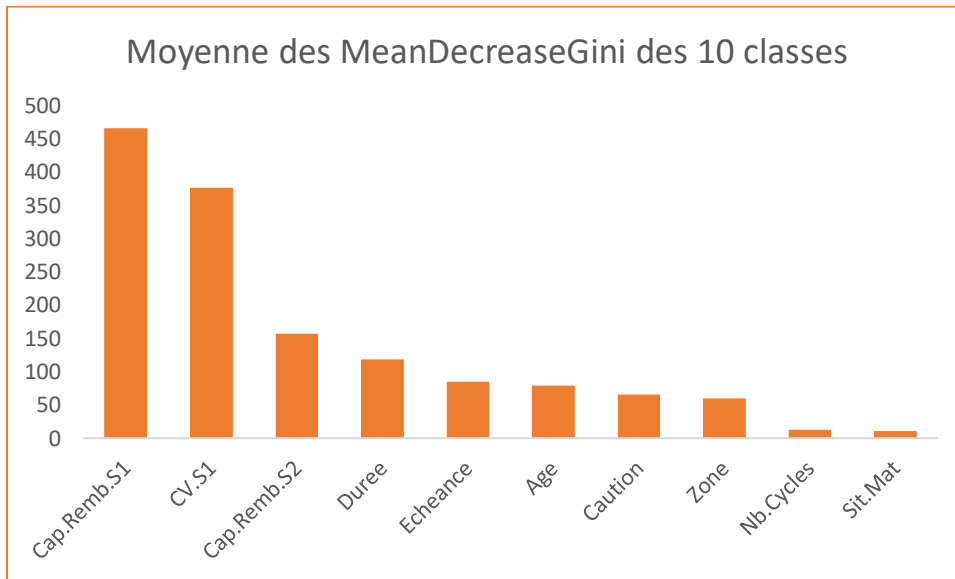


Figure 51 : Importance des variables pour le RF – Cross Validation

Ensuite, on construit les courbes de ROC pour chaque modèle. On obtient :

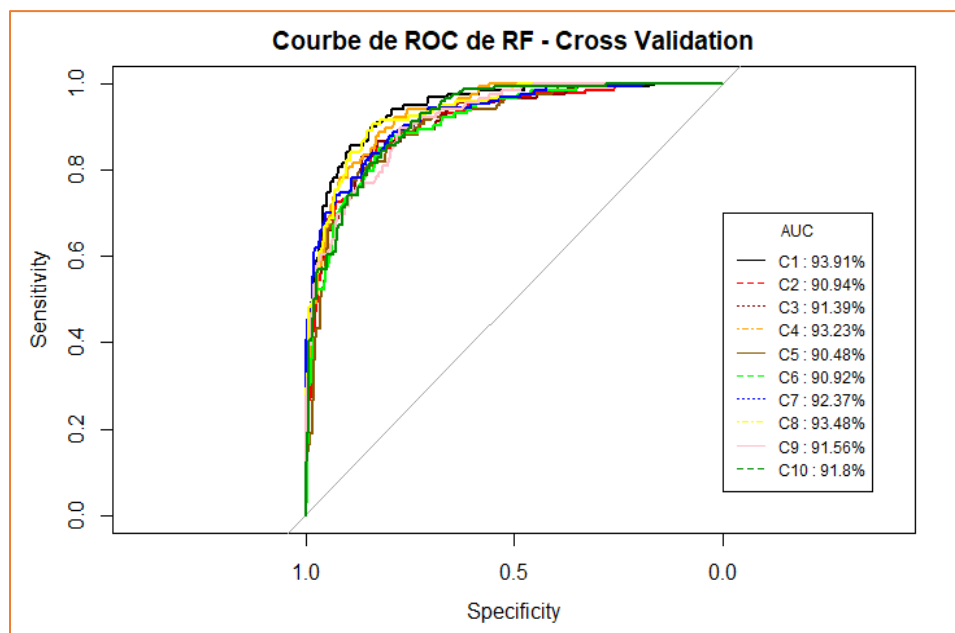


Figure 52 : Courbes de ROC de RF sur les bases test - Cross Validation 10 classes

<sup>14</sup> Voir Annexe : Cross Validation – Random Forest

On remarque clairement que la dispersion de des courbes a diminué. Les surface sous les courbes varient entre 91% et 94% avec une moyenne de 91.98%.

Le modèle du Random Forest diminue la dispersion et améliore la qualité d'ajustement. Ensuite, on mesure la qualité de performance du modèle à l'aide des matrices de confusion.

Modèle	Bonnes prédictions	Sensitivité	Spécificité
Classe 1	87.92%	78.15%	93.40%
Classe 2	84.87%	70.87%	92.61%
Classe 3	84.48%	74.11%	89.41%
Classe 4	86.59%	78.07%	90.83%
Classe 5	84.82%	75.86%	88.93%
Classe 6	83.88%	74.34	88.74%
Classe 7	84.74%	76.42%	88.93%
Classe 8	86.67%	76.27%	91.74%
Classe 8	84.29%	73.28%	89.74%
Classe 9	82.98%	73.97%	88.56%
Moyenne	85.12%	75.12%	90.28%

Tableau 17 : Indicateurs de la performance du RF - Cross Validation 10 Classes

Le taux des bonnes prédictions varie entre 84% et 88% avec une moyenne de 85%, la sensibilité varie entre 74% et 78% avec une moyenne de 75%, finalement, la spécificité varie entre 88% et 93% avec une moyenne de 90%.

On constate que ces taux en moyen, sont très proches aux taux trouvés dans le premier modèle.

#### 4.3.4. Gradient Boosting

Pour le Gradient Boosting, on construit des modèles de 10000 arbres avec un taux d'apprentissage de 0.01. Les graphes<sup>15</sup> du choix du nombre optimal des arbres montre qu'en moyen, le nombre optimal d'arbres est d'environ 4000.

Ensuite, on calcule l'influence relative des variables sur chaque modèle. On constate que l'ordre d'importance des variables est le même pour tous les modèles.

<sup>15</sup> Voir Annexe : Cross Validation – Gradient Boosting

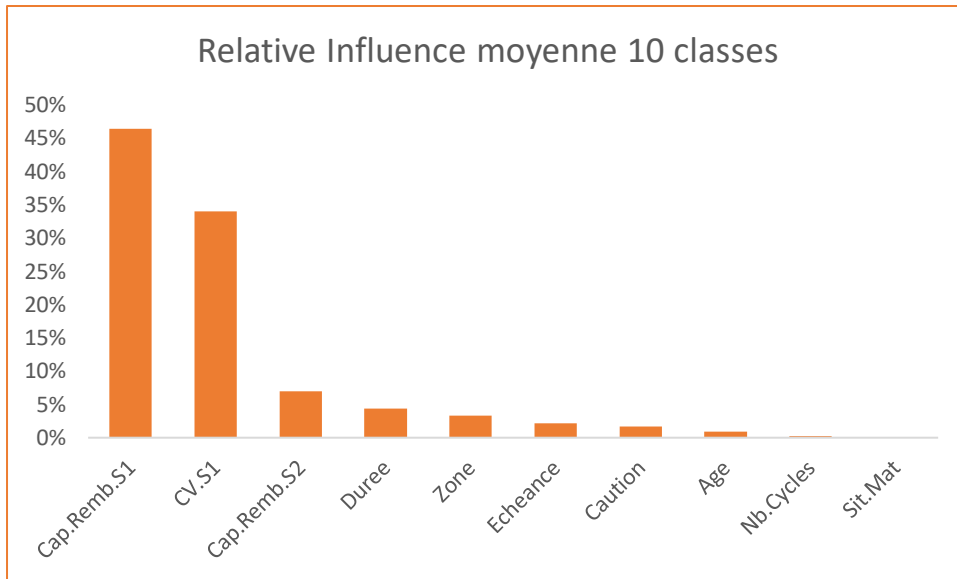


Figure 53 : Influence des variables sur le GBM - Cross Validation

D'après ce graphe, les variables qui ont plus d'influence sur les modèles sont Cap.Remb.S1 et CV.S1.

Ensuite, on trace les courbes de ROC des 10 modèles :

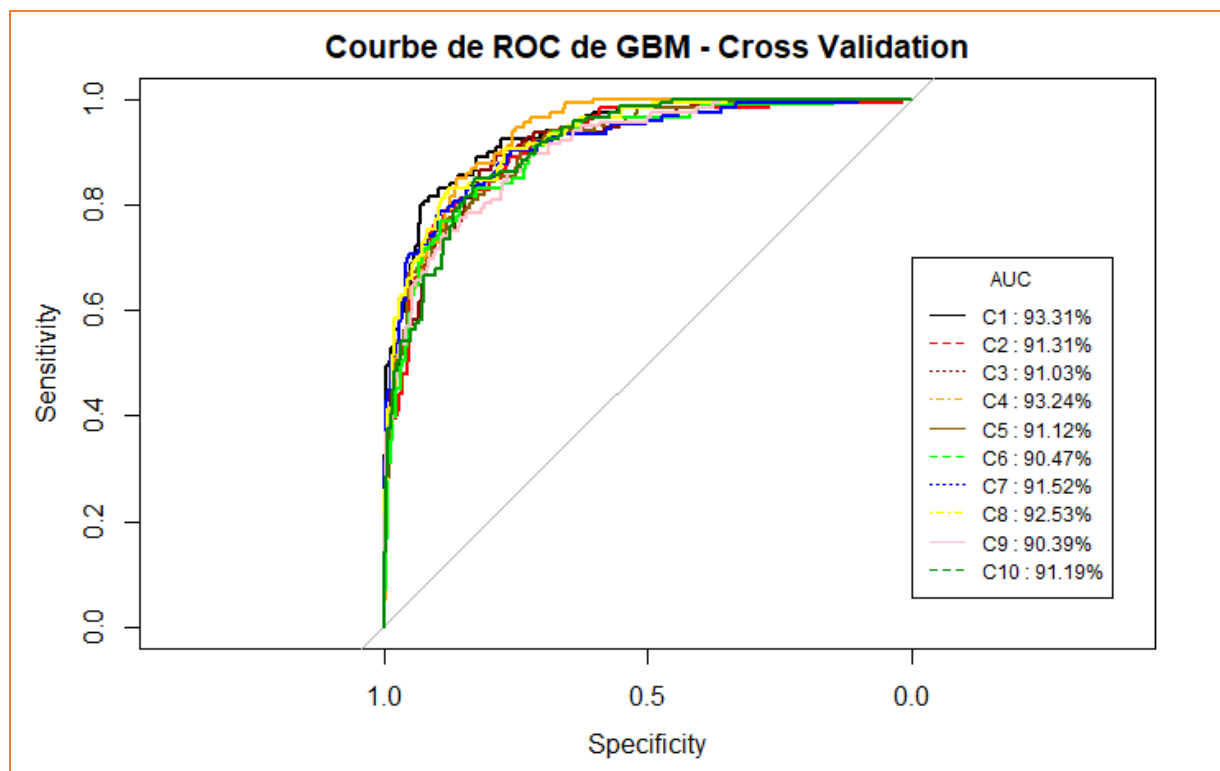


Figure 54 : Courbes de ROC de GBM sur les bases test - Cross Validation 10 classes

Graphiquement, on constate que les courbes sont très proches, En termes de surfaces sous la courbe, elles varient entre 91% et 93% avec une moyenne de 91.58%.

Comme dans le cas du RG, le modèle GBM réduit la dispersion des modèles et améliore la qualité d'ajustement.

Finalement, on mesure la performance de ce modèle :

Modèle	Probabilité	Bonnes prédictions	Sensitivité	Spécificité
Classe 1	0.5	86.36%	76.41%	91.49%
	0.38	85.23%	84.80%	85.46%
Classe 2	0.5	86.39%	76.16%	91.66%
	0.38	85.23%	84.97%	85.35%
Classe 3	0.5	87.77%	75.91%	90.85%
	0.38	84.42%	84.88%	84.17%
Classe 4	0.5	86.45%	77.08%	92.27%
	0.38	85.01%	85.22%	84.90%
Classe 5	0.5	86.59%	77.16%	91.45%
	0.38	84.92%	85.47%	84.64%
Classe 6	0.5	86.11%	76.58%	91.02%
	0.38	84.75%	84.72%	84.77%
Classe 7	0.5	86.59%	77.41%	91.32%
	0.38	85.35%	85.38%	85.33%
Classe 8	0.5	85.94%	75.91%	91.10%
	0.38	85.12%	85.05%	85.16%
Classe 9	0.5	86.08%	76.58%	90.98%
	0.38	84.70%	84.88%	84.60%
Classe 10	0.5	86.22%	76.00%	91.49%
	0.38	85.15%	84.63%	85.61%
Moyenne	0.5	86.25%	76.52%	91.26%
	0.38	84.99%	85.00%	84.98%

Tableau 18 : Indicateurs de la performance du GBM - Cross Validation 10 Classes

La première probabilité 0.5 correspond au seuil qui maximise le taux de bonnes prédictions, la deuxième correspond au seuil d'intersection des trois 3. Et cela appliqué sur un vecteur des probabilités moyennes des 10 modèles.

La première remarque concerne les seuils. D'une part, ils vérifient les conditions sur les 10 modèles. D'autre part, ils sont presque égaux avec le premier modèle trouvé auparavant (même seuil d'intersection 0.38, et une petite différence entre les seuils qui minimisent l'erreur de prédiction : 0.48 contre 0.5).

En termes de performance, on constate que les résultats sont également très proches.

#### 4.3.5. Comparaison des modèles

En premier lieu, les quatre modèles confirment l'importance de deux variables : Cap.Remb.S1 et CV.S1.

Ensuite, pour comparer les quatre modèles, on utilise des modèles moyens entre les 10 modèles tels que :

- Pour la régression logistique, le modèle moyen correspond au modèle dont les estimations de ses paramètres égalent les moyennes des estimations sur les 10 modèles.
- Pour l'arbre de classification, le modèle moyen correspond au nombre total des votes sur les 10 arbres.
- De même pour le RF, on agrège en moyennant les votes.
- Finalement, pour le GBM, le modèle moyen correspond aux probabilités moyennes sur les 10 modèles.

Ensuite, on trace les courbes de ROC pour ces modèles :

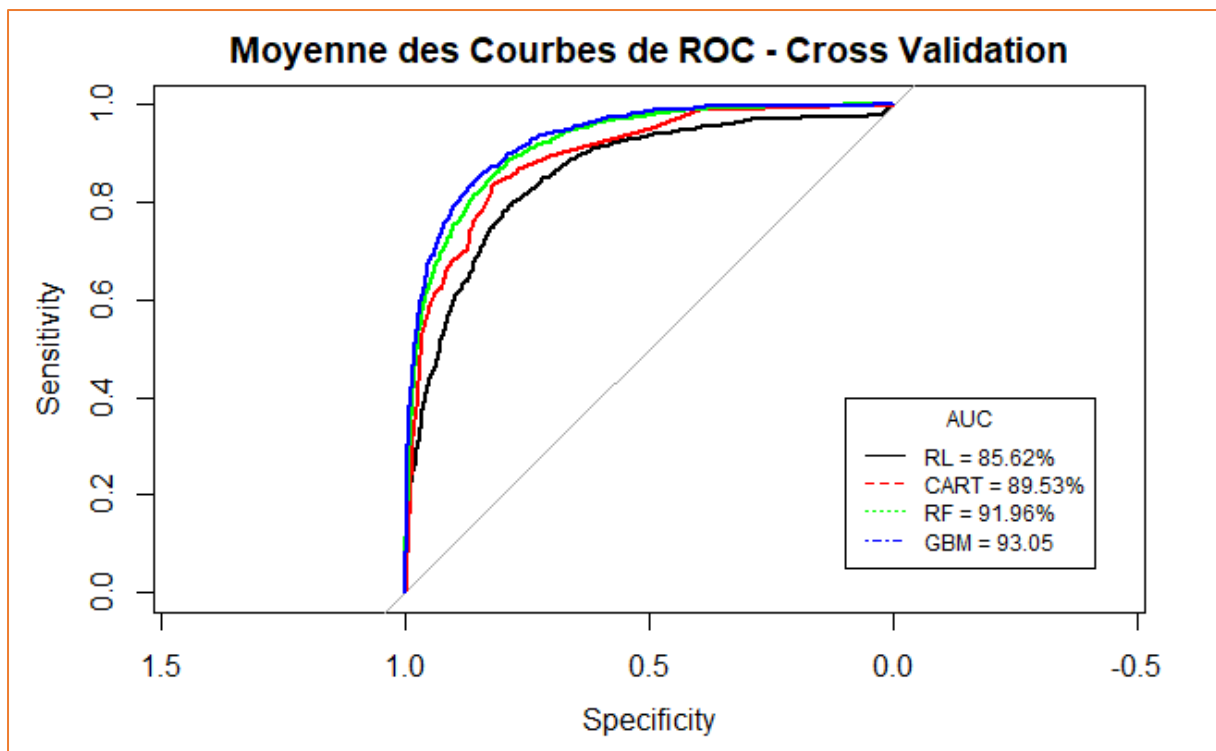


Figure 55 : Comparaison des courbes de ROC entre les 4 modèles - Cross Validation

Encore, les modèles d'apprentissage améliorent la qualité d'ajustement des données, et en particulier les modèles d'agrégation. En plus de cela, ces modèles réduisent la sensibilité à la base d'entraînement, ainsi que la performance de prédictions.

Ces modèles sont par construction, établis en combinant des arbres construits sur des échantillons tirés aléatoirement, donc par défaut, ils sont construits avec une technique semblable à la Cross Validation.

La régression logistique et le Gradient Boosting permettent d'estimer les probabilités de défaut en premier lieu, ce qui nous donne la possibilité de choisir les seuils de détermination du défaut selon les besoins et les priorités.

On peut faire la même chose pour le modèle du RF en cas de besoin. A partir des probabilités obtenues par les votes, on pourrait choisir des seuils comme dans le cas des autres modèles. En utilisant le modèle moyen du RF, on trace les taux d'erreurs de prédictions en fonction des probabilités :

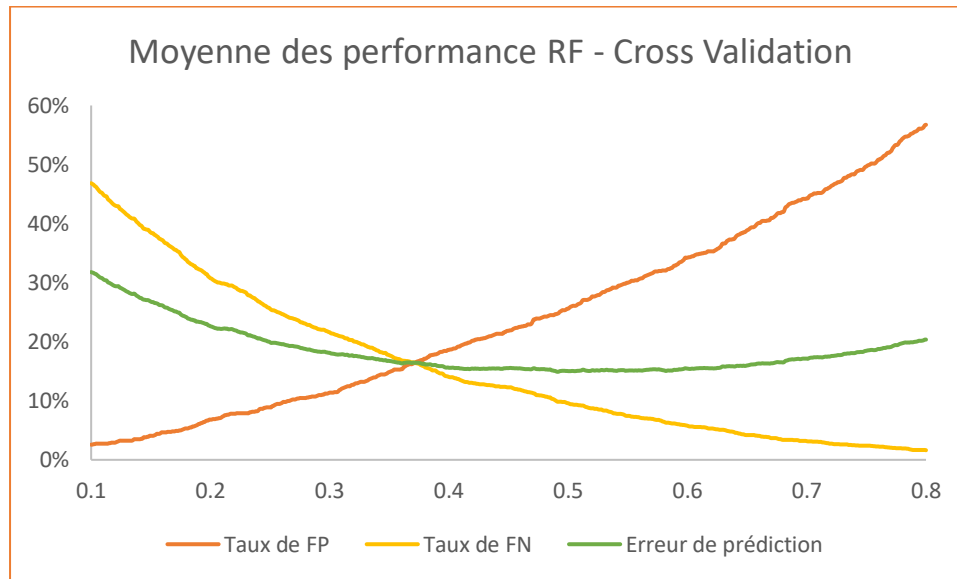


Figure 56 : Taux de performance du RF en fonction des probabilités – Cross Validation

Les courbes des taux de performance sont similaires aux courbes du modèle GBM, le seuil qui minimise le taux d'erreur de prédiction correspond également à 0.5, et la probabilité d'intersection des 3 courbes est égale 0.37 avec une erreur égale à 16%, ce qui est également très proche à la performance du GBM (seuil = 0.38 et erreur = 15%).

## 4.4. Choix du modèle

### 4.4.1. Choix du modèle

Les modèles d'agrégation par apprentissage automatique, RF et GBM, ajustent bien les données, donnent de bonnes prédictions sur les trois et assurent la stabilité par rapport au choix de la base d'apprentissage.

La limite de ces deux méthodes se trouve dans leurs complexités qu'on peut résumer dans les points suivants :

- La difficulté d'interprétation du modèle obtenu : les algorithmes s'exécutent en boîte noire. Même si on arrive à représenter les 1000 arbres, on ne pourrait pas définir la tendance de variation de la probabilité de défaut en fonction des

variables explicatives.

- Malgré la similarité et la stabilité des résultats, les paramètres de tirage aléatoire des échantillons génèrent des modèles différents dans chaque exécution.
- Nécessité d'un logiciel dédié à la construction de ce type de modèles, ainsi que des experts. Cela pose des difficultés et des coûts supplémentaires pour implémentation de ce type de modèle dans le système d'information de l'institution.

Pour ces raisons on choisit le modèle de la régression logistique, ce modèle est à la fois simple, utilise une formule qui facilite l'interprétation du sens de variations de la probabilité de défaut, ainsi il permet aux décideurs de choisir les seuils convenables à leur stratégie commerciale.

Finalement, le modèle choisi du modèle de la régression logistique correspond au modèle moyen établi par la Cross Validation :

$$\begin{aligned} \text{logit}(p(x)) = & 0.134 - 0.002 \times \text{Age} + 0.062 \times 1_{\text{Zone2}} + 0.0.133 \times 1_{\text{Zone3}} + 0.223 \\ & \times 1_{\text{Zone4}} - 0.037 \times 1_{\text{PremierCrédit}} + 0.001 \times \text{Duree} - 0.034 \times \text{Caution} \\ & - 0.138 \times \text{Cap. Remb. S1} + 0.142 \times \text{CV. S1} - 0.061 \times \text{Cap. Remb. S2} \end{aligned}$$

#### 4.4.2. Grille de notation

A partir du modèle précédent, on trace les taux de performance de prédiction du modèle en fonction de probabilités :

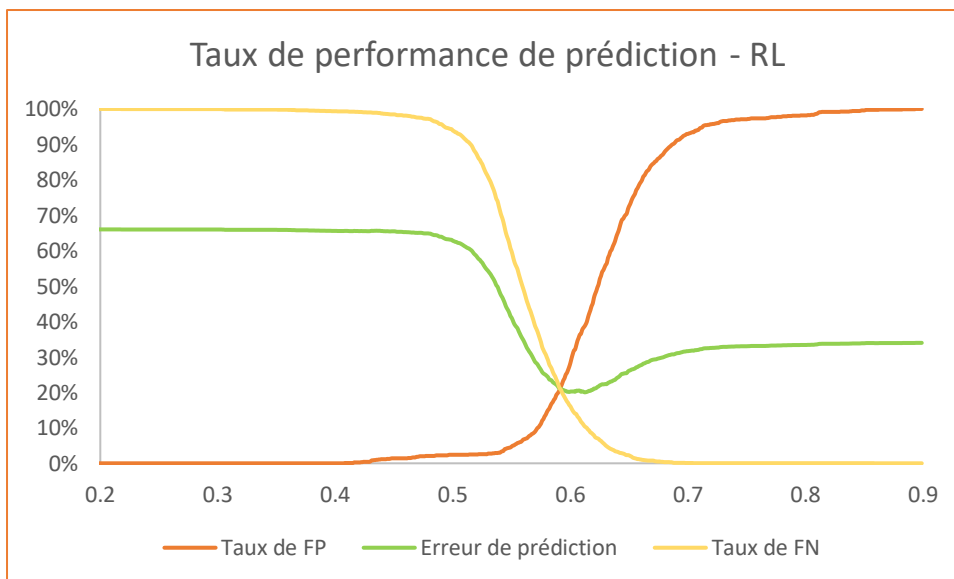


Figure 57 :: Taux de performance de la RL en fonction des probabilités – Cross Validation

L'objectif de cette partie consiste à trouver des seuils pour déterminer des classes de risque selon la probabilité de défaut. Cette classification aiderait les agents de l'institution à prendre la décision d'autoriser ou de rejeter une nouvelle demande.

Pour cela, on propose 4 classes :

- Classe verte : risque faible
- Classe jaune : risque moyen
- Classe orange : risque élevé
- Classe rouge : risque très élevé

Pour la classe verte, on choisit un seuil de 0.552. Les probabilités inférieures à ce seuil se caractérisent par sensibilité minimale de 95%. Autrement dit, le risque pour qu'un individu qui a eu réellement un défaut et ne soit pas détecté est inférieur à 5%.

Pour la classe jaune, on choisit le seuil d'intersection des 3 courbes : 0.591. Dans ce seuil, les 3 erreurs sont égales à 22%. Par conséquent, le risque pour qu'un individu qui a eu réellement un défaut et ne soit pas détecté est inférieur à 22%.

Pour la classe orange, on choisit un seuil de 0.623. A partir de ce seuil, la sensibilité décroît au-dessous de 50%. C'est-à-dire que pour cette classe, on est déjà sur une région où la probabilité de déclarer un défaut incorrectement est inférieure à 0.5.

La classe rouge est le complémentaire des autres groupes, les chances d'avoir un défaut sont très élevés.

Les décisions à prendre selon ces classes :

- Classe verte : Accorder les crédits sans trop de contraintes.
- Classe Jaune : Demander plus de cautions.
- Classe orange : Traiter les dossiers un par un par l'agent ou le comité des crédits
- Classe rouge : Rejeter les crédits.

# Conclusion

A cause de la faible qualité des données dans la microfinance à cause de l'asymétrie d'information d'une part et de la mauvaise collecte par les IMF d'autre part, toute modélisation statistique et notamment le Credit Scoring, constitue une tâche difficile. Elle nécessite des traitements spéciaux des variables afin de créer les variables nécessaires.

La comparaison entre les quatre modèles : régression logistique, arbre de classification CART, forêt aléatoire et le Gradient Boosting, indique que les méthodes d'apprentissage automatique et surtout ceux basées sur l'agrégation sont les plus performantes.

A cause de leurs complexités et de difficultés à les implémenter dans les systèmes d'information de l'IMF, on choisit le modèle de la régression logistique, à partir duquel on construit une grille de notation des clients selon leurs niveaux de risque.

## **Bibliographie**

- [1] Comité de Bâle sur le contrôle bancaire, Vue d'ensemble du Nouvel accord de Bâle sur les fonds propres, Avril 2003
- [2] Vers un monde sans pauvreté, Muhammad Yunus 1998
- [3] Leo Breiman - Classification and Regression Trees-CRC
- [4] Leo Breiman Random Forest January 2001
- [5] H. Friedman The Element of statistical learning Second Edition 2009
- [6] P. McCULLAH Generalized Linear Models

### **Support de cours :**

- [7] CHAOUBI A. Modèles Linéaires Généralisés 2019-2020
- [8] ABDELKHALEK T. Micro et Macro économétrie 2019-2020
- [9] BADAOUI F. Analyse des données discrètes 2019-2020
- [10] CHATER M. Economie bancaire et monétaire 2019-2020



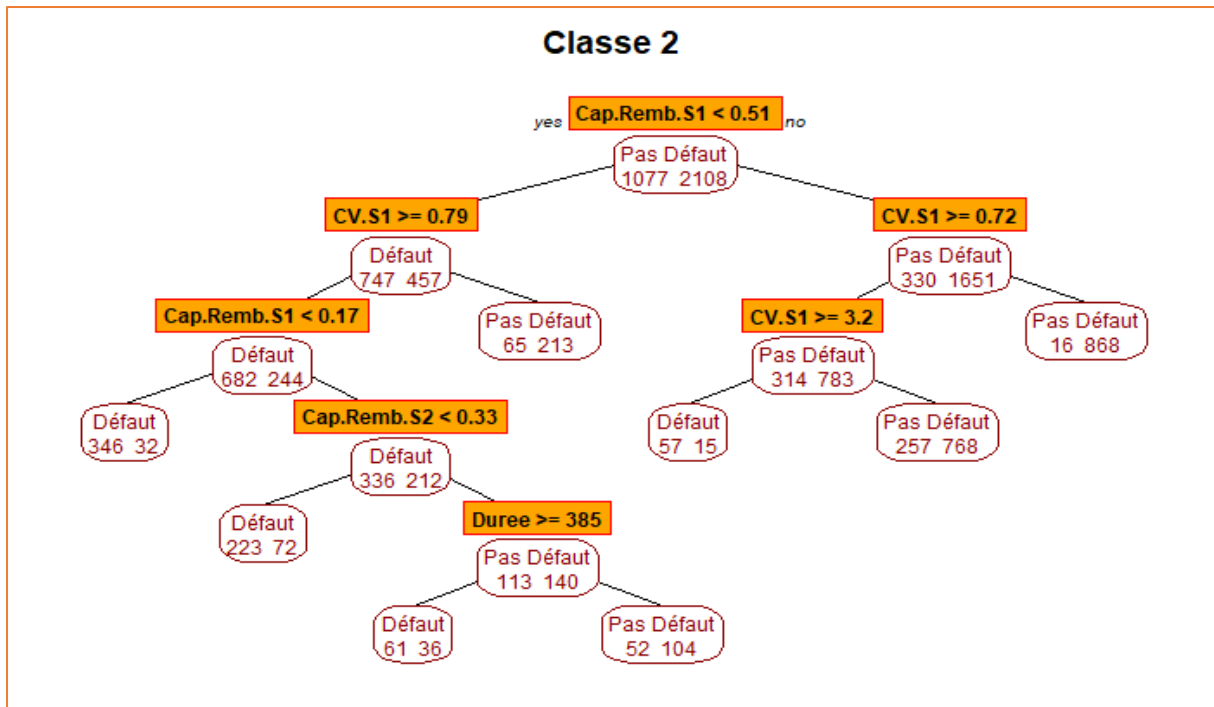


Figure 59 : Arbre de Classification 2 - Cross Validation

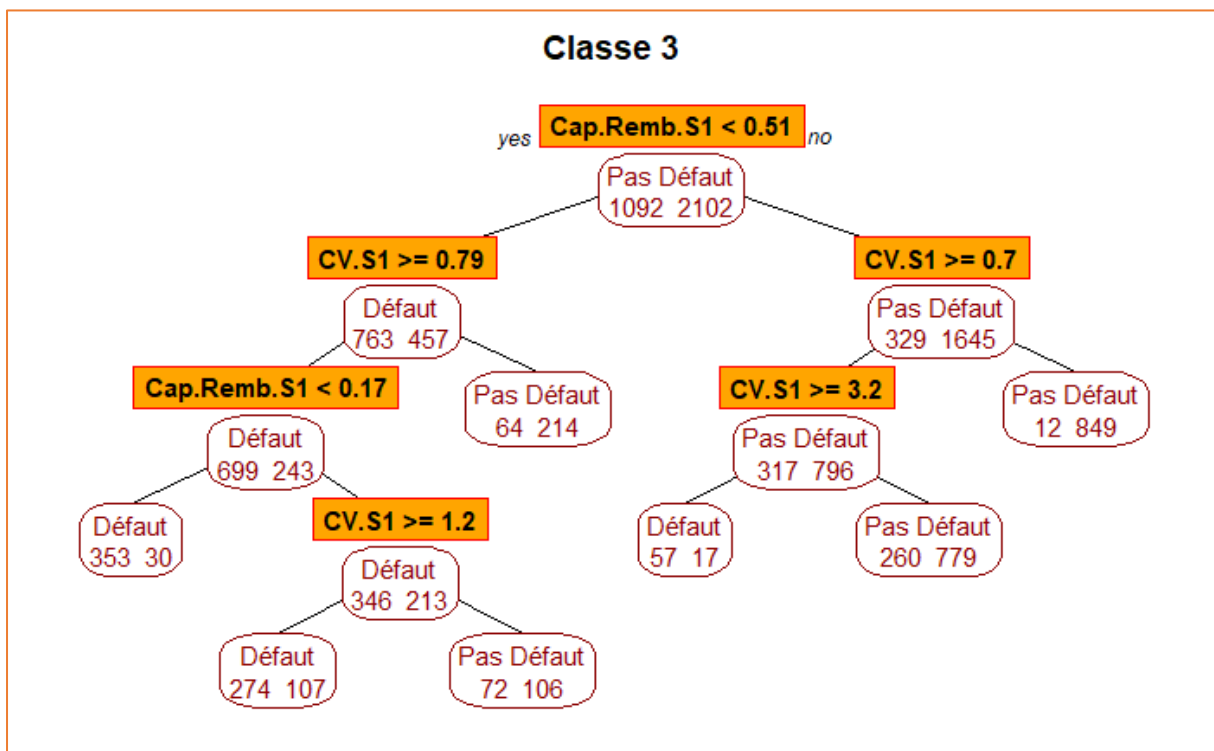


Figure 60 : Arbre de Classification 3 - Cross Validation



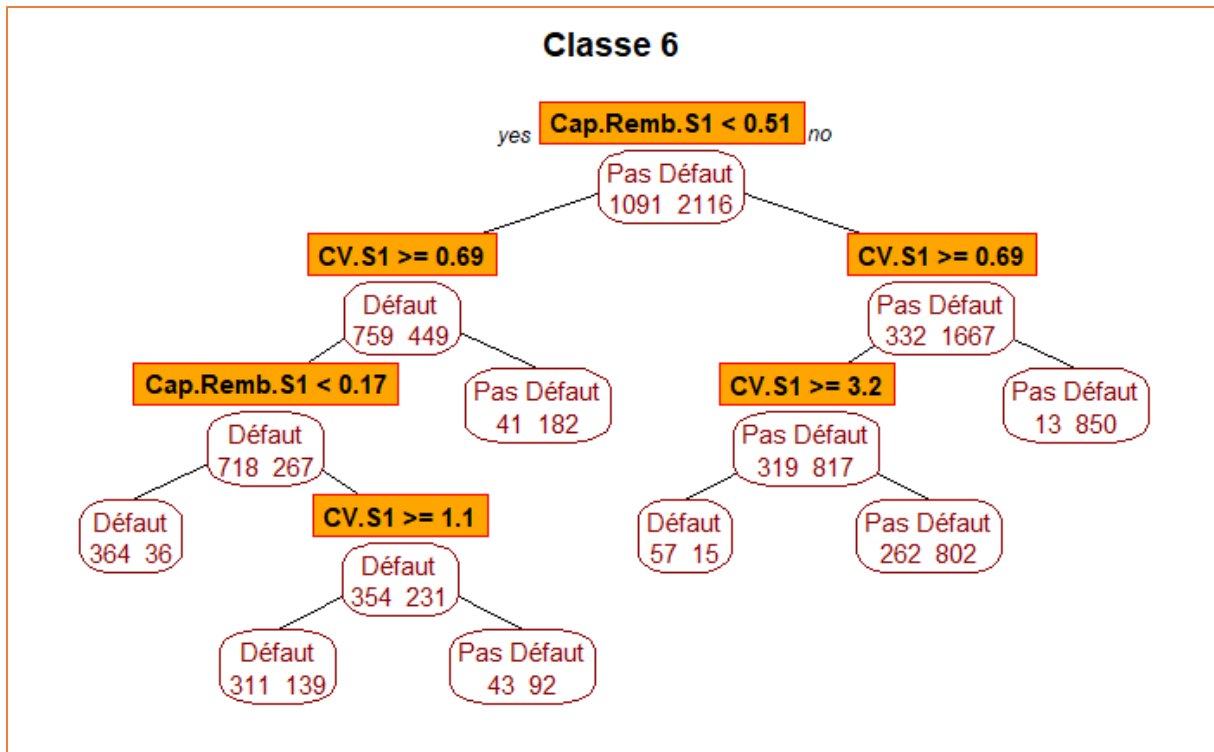


Figure 63 : Arbre de Classification 6 - Cross Validation

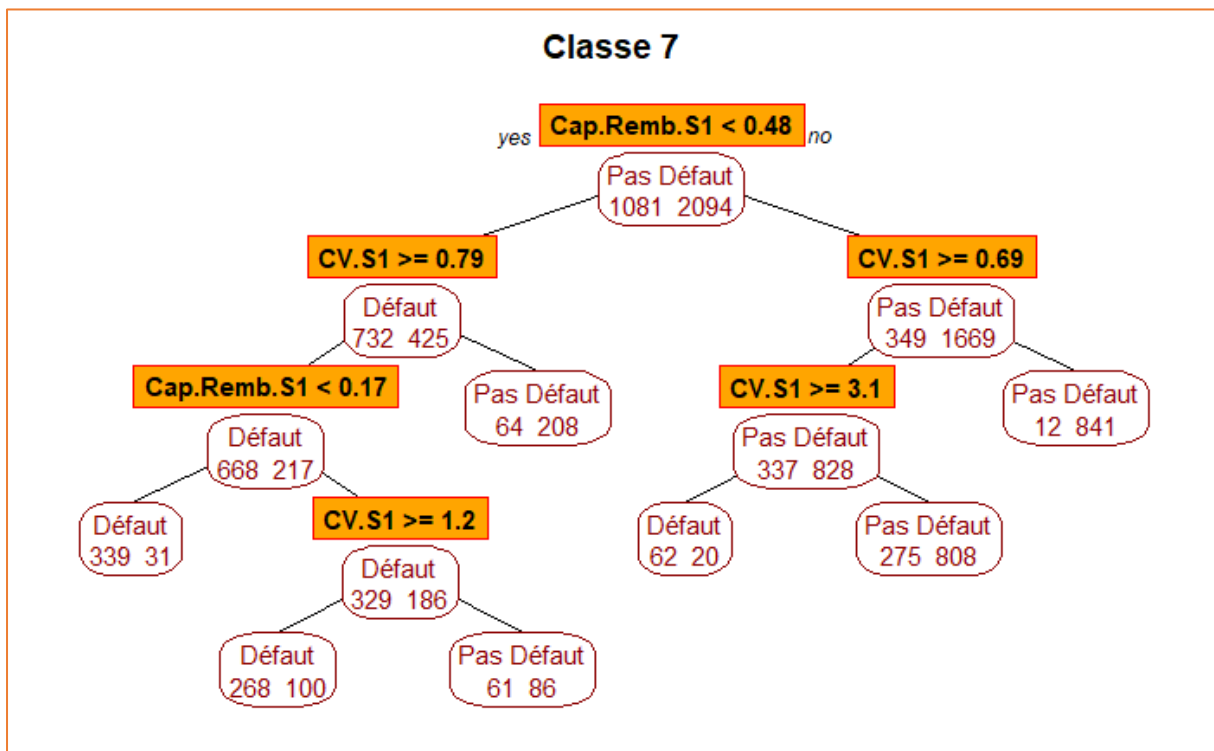


Figure 64 : Arbre de Classification 7 - Cross Validation

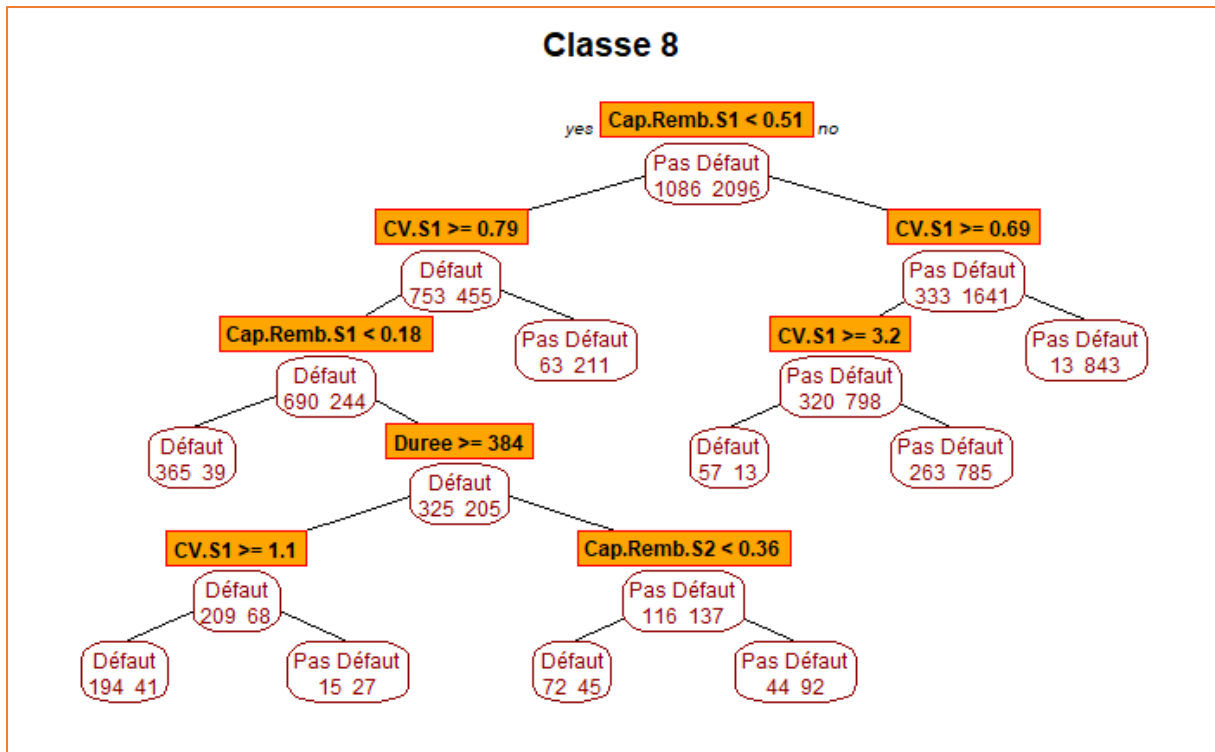


Figure 65 : Arbre de Classification 8 - Cross Validation

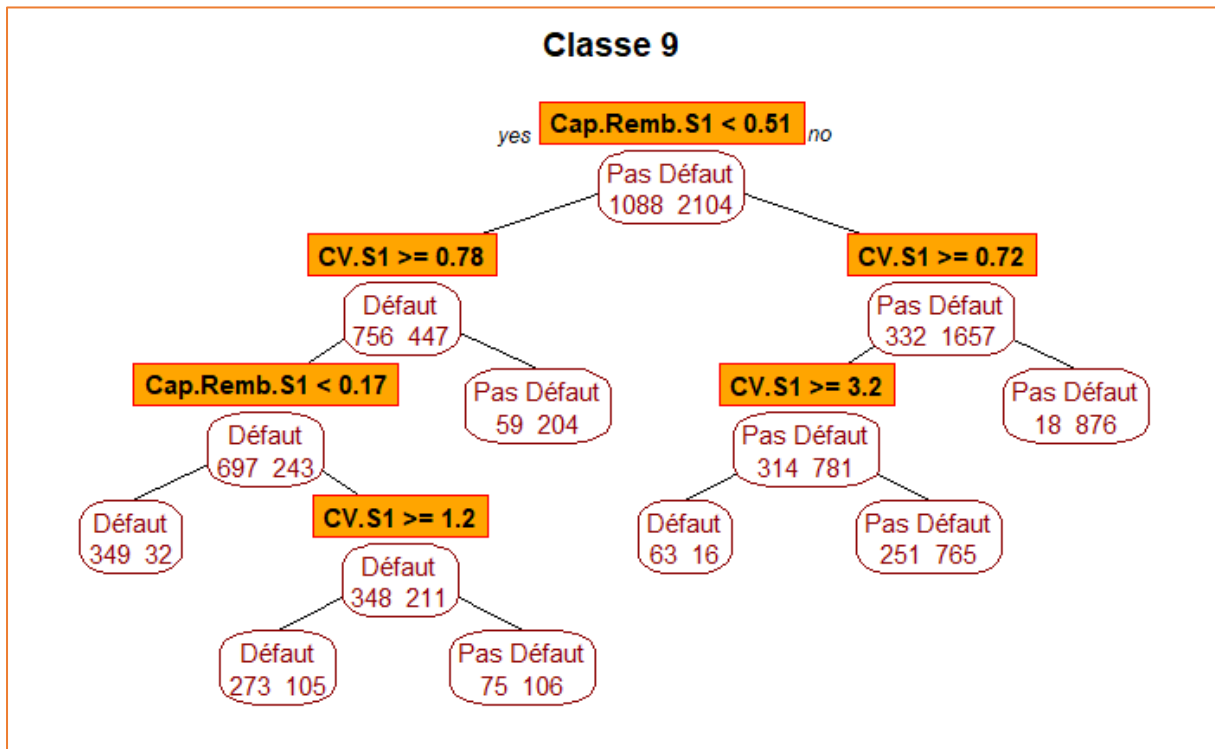
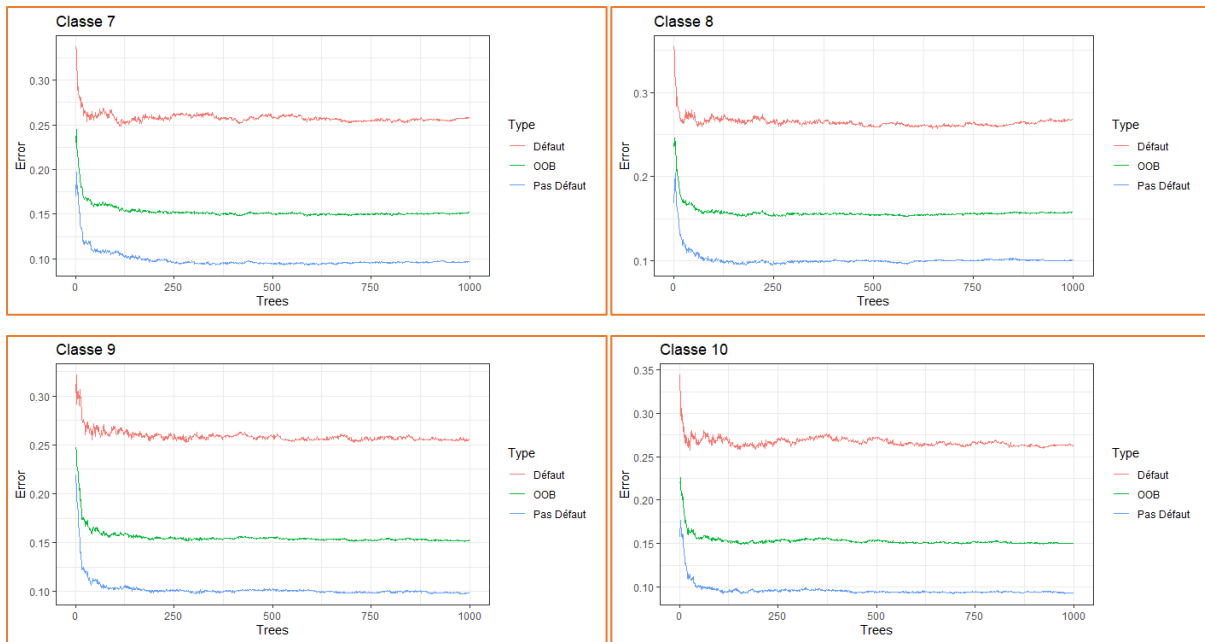


Figure 66 : Arbre de Classification 9 - Cross Validation





### 2.3.Gradient Boosting

