



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

Conception d'un modèle de notation pour l'octroi de microcrédit

Préparé par : *Mme. Essi Valentine MAMATTAH*

Sous la direction de : *M. Abdellah MANADIR (INSEA)*
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. Abdellah MANADIR (INSEA)*
- *M. Fouad MARRI (INSEA)*
- *M. Hamza BENFDIL (BCP CONSULTING)*

Résumé :

L'activité bancaire n'est pas une activité comme les autres en raison, d'une part, des risques spécifiques qu'elle fait courir à la collectivité : perte de l'épargne des déposants, crise systémique en cas de défaillance d'un ou plusieurs établissements de crédit voire tout le système bancaire ; et d'autre part, le risque est la principale source de profits pour une banque. Ainsi la banque doit bien évidemment prendre des risques, mais ces derniers doivent être évalués ou mesurés pour éviter des pertes considérables. L'un des principaux risques et non des moindres auxquels une banque fait face reste sans doute le *risque de crédit* qui résulte de l'incertitude quant à la possibilité ou la volonté des contreparties ou des clients à remplir leurs obligations vis-à-vis de la banque.

Ce travail a pour principal but de concevoir un modèle capable de situer un client qui aspire à un crédit quant à sa probabilité à honorer ses engagements vis-à-vis de la banque. Pour ce faire, la modélisation fera intervenir plusieurs variables susceptibles d'influencer la probabilité de défaut d'un emprunteur et ne gardera que celles qui seront les plus significatives. Afin d'avoir le modèle le plus parcimonieux possible, deux méthodes seront utilisées : la régression logistique (qui est la plus évidente, la plus utilisée et la plus connue) et une méthode de machine learning dite de Random Forest.

L'élaboration de ce modèle nous permettra par la suite d'établir une grille de notation qui affectera chaque client à une classe et qui donnera en fonction de la classe où il se situe la décision à prendre quant à l'acceptation de la demande d'octroi de crédit, au refus ou encore à savoir si ce client mérite une attention particulière ou une exploration d'autres informations additionnelles.

Sur 16 variables de départ, seules 6 ont été retenues dans la conception du modèle dont la plus importante et la plus significative est la « Capacité de remboursement du client ».

Mots clés :

Scoring, probabilité de défaut, risque de crédit, modèle logistique, Bâle II.

Dédicaces :

À la mémoire de mon père, ma motivation, tu restes à tout jamais dans mon cœur.

À mon adorable mère dont la force et le courage m'impressionnent toujours.

À ma sœur, mes frères, cousins et cousines pour leur soutien.

À toute ma famille.

À tous mes amis.

À tous ceux et celles qui ont apporté leurs contributions à la réalisation de ce

travail.

Remerciements :

Mes remerciements, vont, en premier lieu, à l'endroit de mon encadrant au sein de BCP Consulting, M. Hamza BENFDIL qui m'a conduit et m'a fait bénéficier de son expérience tout au long de mon stage.

Je tiens à exprimer mes remerciements et ma profonde gratitude à M. Abdellah MANADIR, qui a su avoir confiance en moi et qui a accepté d'être mon tuteur au sein de l'Institut National de Statistique et d'Economie Appliquée. Ses conseils, directives et enseignements m'ont été très favorables à la réalisation de ce travail.

Je remercie également M. Fouad MARRI d'avoir accepté d'évaluer ce travail.

Je ne saurais terminer sans adresser mes sincères remerciements au personnel de BCP Consulting, aux corps administratif et enseignant de l'INSEA, à mes camarades et à tous ceux qui, de près ou de loin, m'ont aidée dans cette stimulante expérience.

Table des matières

Résumé :.....	3
Dédicaces :.....	4
Remerciements :.....	5
Table des matières.....	6
Liste des abréviations.....	10
Liste des tableaux.....	11
Liste des figures.....	11
Introduction générale :.....	13
Chapitre 1 : Contexte général.....	14
I. Présentation de l'organisme d'accueil :.....	14
1. Généralités sur BCP Consulting :.....	14
2. Domaines d'actions :.....	14
3. Quelques indicateurs clés :.....	15
II. Risque bancaire : Risque de crédit.....	17
1. Risque de crédit :.....	17
2. Probabilité de défaut :.....	17
3. Evénement de défaut :.....	18
III. De Bâle I vers Bâle II :.....	18
1. L'approche classique de la gestion des risques : RATIO COOKE (Bâle I) :.....	19
2. L'approche dynamique de la gestion des risques : RATIO MC DONOUGH (Bâle II) :.....	20
3. Approche Quantitative de la mise en place d'un SNI des entreprises telle que recommandée par Bâle II :.....	21

Chapitre 2 : Elaboration d'un modèle de notation interne _____ – Cadre théorique	22
.....	22
I. Notation externe Vs notation interne :.....	22
1. Notation externe :.....	22
2. Notation interne :	24
II. Méthodologie d'élaboration d'un modèle de score :	25
1. Architecture générale d'un système de notation interne :	26
2. Le modèle de crédit scoring :.....	28
2.1.Principes généraux du crédit scoring :.....	28
2.2.Les types de modèles de crédit scoring :	29
2.2.1. Les modèles déductifs ou à priori :	29
2.2.2. Les modèles empiriques ou basés sur l'historique :.....	30
2.3.Démarche de construction d'un modèle de crédit scoring :	30
2.3.1. Collection des données :.....	31
2.3.2. Spécification des données :.....	32
2.3.3. Traitement des valeurs manquantes et des valeurs aberrantes :	33
2.3.4. Sélection des variables explicatives :.....	34
2.3.5. Choix de la méthode statistique :	35
2.3.6. Modélisation et tests :.....	36
2.3.7. La validation :.....	36
2.3.8. Décision du besoin d'ajustement :	37
III. Techniques de classification et de validation du modèle de scoring : 38	
1. La régression logistique :	38
1.1.Les hypothèses de la régression logistique :.....	40
1.2.Méthodes de sélection des variables :.....	41

1.2.1. La forward Selection (FS) :	41
1.2.2. La Backward Elimination (BE) :	42
1.2.3. La Méthode Mixte (STEPWISE) :	42
1.3. Validation et évaluation du modèle :	42
1.3.1. Validation :	43
1.3.1.1. Test du rapport de vraisemblance :	43
1.3.1.2. Test de stabilité :	44
1.3.1.3. Mesure AIC et BIC :	44
1.3.1.4. Test de significativité :	45
1.3.2. Evaluation :	46
1.3.2.1. Le test de performance :	47
1.3.2.2. La matrice de confusion :	47
1.3.2.3. La ROC curve :	48
1.4. Adéquation du modèle :	49
2. Famille de modèles aléatoires : Bagging et Random Forest (les forêts aléatoires) :	51
2.1. Bagging : Principe et algorithme	51
2.2. Forêts aléatoires :	54
2.3. Les critères d'évaluation :	57
3. Comparaison des deux techniques de rating :	58
Chapitre 3 : Elaboration d'un modèle de notation interne – Partie Pratique	
.....	60
I. Données et statistique descriptive des variables :	60
1. Analyse de la base de données :	60
2. Traitement de la base et analyse descriptive des données :	62
2.1. Valeurs manquantes :	62

2.2. Valeurs aberrantes :	65
2.3. Répartition des clients sains et en défaut :	66
2.4. Analyse des corrélations :	67
II. Elaboration du modèle :	70
1. Régression logistique :	70
1.1. Sélection des variables :	70
1.1.1. Par la méthode backward :	70
1.1.2. Par la méthode forward :	74
1.1.3. Par la méthode stepwise :	75
1.2. Comparaison des 3 modèles issus de la régression logistique : ...	76
2. Random Forest :	77
2.1. Nombre d'arbres à considérer :	77
2.2. Nombre de variables à considérer :	78
2.3. Importance des variables :	79
2.4. Pouvoir prédictif du modèle :	79
2.5. Validation sur la base de test :	80
3. Choix du modèle final pour la grille de notation :	82
4. Grille de notation :	84
III. Implémentation :	86
1. Notation individuelle :	86
1. Notation par groupe de clients :	88
Conclusion générale	90
Bibliographie	91
Annexes	92

Liste des abréviations

INSEA : Institut National de Statistique et d'Economie Appliquée

BCP : Banque Centrale Populaire

PD : Probabilité de Défaut

AIC : Akaike Information Criteria / Critère d'Information d'Akaike

BIC : Bayesien Information Criteria

ROC curve : Receiver Operating Characteristic curve

AUC : Area Under Curve

BIS : Bank for International Settlements / Banque des règlements internationaux

NI : Notation Interne

EL : Expected Loss ou Perte attendue

LGD : Loss Given Default

EAD : Exposure At Default

SNI : Système de notation interne

S&P : Standard and Poor's

ECL : Expected Credit Loss

IRB : Internal Rating Based

FP : Fonds Propres

CART : Classification And Regression Trees

OOB : Out Of Bag

VP : Vrais Positifs

VN : Vrais Négatifs

FP : Faux Positifs

FN : Faux Négatif

Liste des tableaux

<i>Tableau 1 : Les piliers de Bâle II</i>	20
<i>Tableau 2 : Symboles de notation des agences de notation</i>	23
<i>Tableau 3 : Comparaison entre la régression logistique et la méthode Random Forest</i>	59
<i>Tableau 4 : Les variables et leurs types</i>	61
<i>Tableau 5 : Valeurs manquantes</i>	63
<i>Tableau 6 : Récapitulatif des corrélations</i>	69
<i>Tableau 7 : Interprétation AUC en fonction des valeurs prises</i>	73
<i>Tableau 8 : Comparaison des 3 modèles sur la base de la méthode de sélection</i>	76
<i>Tableau 9 : Prédiction sur base de test / Random Forest</i>	81
<i>Tableau 10 : Performance des 2 modèles</i>	82
<i>Tableau 11 : Grille de notation</i>	84
<i>Tableau 12 : Décision selon la classe d'appartenance</i>	86

Liste des figures

<i>Figure 1 : Pays d'intervention de BCP Consulting</i>	15
<i>Figure 2 : Approches de notation interne et externe</i>	25
<i>Figure 3 : Etapes de la conception d'un modèle de credit scoring</i>	31
<i>Figure 4 : Les 3 phases du processus de credit scoring</i>	37
<i>Figure 5 : ROC Curve</i>	49

<i>Figure 6 : Algorithme Bagging</i>	52
<i>Figure 7 : Algorithme Random Forest</i>	55
<i>Figure 8 : Illustration Random Forest</i>	56
<i>Figure 9 : Matrice de confusion</i>	57
<i>Figure 10 : La répartition des variables qualitatives</i>	61
<i>Figure 11 : La distribution des variables quantitatives</i>	62
<i>Figure 12 : Technique de traitement des missing values</i>	64
<i>Figure 13 : Boxplots des variables qualitatives</i>	66
<i>Figure 14 : Effectifs des clients sains et en défaut</i>	67
<i>Figure 15 : Corrélation entre les variables quantitatives</i>	67
<i>Figure 16 : Corrélation variables quantitatives Vs variables qualitatives</i>	68
<i>Figure 17 : Test de nullité globale</i>	70
<i>Figure 18 : Estimation des paramètres du modèle issu de la régression logistique</i>	71
<i>Figure 19 : Odds ratio</i>	71
<i>Figure 20 : Récapitulatif de la méthode de sélection Backward</i>	72
<i>Figure 21 : Courbe ROC Backward</i>	72
<i>Figure 22 : Test de Hosmer Lemeshow</i>	73
<i>Figure 23 : Concordance / Discordance (Régression logistique)</i>	74
<i>Figure 24 : Récapitulatif Forward</i>	74
<i>Figure 25 : Courbe ROC Forward</i>	75
<i>Figure 26 : Récapitulatif Stepwise</i>	75
<i>Figure 27 : Courbe ROC Stepwise</i>	76
<i>Figure 28 : Taux de mauvaises classifications en fonction du nombre d'arbres de la forêt aléatoire</i>	77
<i>Figure 29 : Variation de l'erreur OOB en fonction du nombre de variables à considérer</i>	78
<i>Figure 30 : Importance des variables pour la méthode RF</i>	79
<i>Figure 31 : Recherche du seuil de défaut</i>	83
<i>Figure 32 : Répartition des clients sains en fonction de la note</i>	85
<i>Figure 33 : Répartition des clients en défaut en fonction de la note</i>	85
<i>Figure 34 : Notation individuelle</i>	87
<i>Figure 35 : Sortie Scoring individuel</i>	88
<i>Figure 36 : Notation collective Scoring</i>	89
<i>Figure 37 : Sortie Scoring par groupe de clients</i>	89

Introduction générale :

Les établissements bancaires ont pour priorité l'anticipation des risques qui se rapportent aux crédits. Cette analyse permet d'identifier les potentiels risques avant qu'ils ne se produisent. Il existe de nombreuses techniques permettant de quantifier et d'évaluer les dangers de chaque portefeuille. La banque gagne à la fois en temps et en argent à écarter les risques au sein de sa clientèle. En effet dès qu'un risque apparaît, il faut rapidement le gérer, ce qui mobilise des moyens humains mais aussi financiers. Lorsque la situation du client se dégrade, la banque n'est jamais totalement sûre de récupérer l'intégralité de son investissement. C'est cette incertitude constante qui fait peur aux banques. L'anticipation par l'analyse en amont des risques permet de combler une partie de cet avenir incertain et de sécuriser l'activité de crédit.

Les établissements bancaires sont donc dans l'obligation de prendre des sécurités pour garantir les engagements. En effet les risques liés aux crédits sont nombreux et la situation de l'emprunteur peut rapidement se dégrader. Avec ses techniques les banques augmentent leurs chances d'obtenir un remboursement total du prêt et dans les temps. Il est compréhensible qu'aucun organisme ne prête des fonds à un tiers sans avoir un minimum de sécurité pour palier des événements inattendus. Ainsi, les banques trouvent des solutions pour gérer correctement le risque de contrepartie pour ne pas engager directement une gestion curative souvent longue et coûteuse.

C'est dans cette perspective que s'inscrit cette étude qui vise à concevoir un modèle afin de prévoir la probabilité de défaut d'un client avant l'octroi de microcrédit.

Chapitre 1 : Contexte général

I. Présentation de l'organisme d'accueil :

1. Généralités sur BCP Consulting :



BCP Consulting est une structure de conseil dédiée à l'accompagnement au développement à l'international du Groupe BCP. Elle a été créée afin d'accélérer la capacité d'exécution des orientations stratégiques et de faciliter la conduite des projets majeurs initiés, quelle que soit leur nature, tout en capitalisant sur les synergies potentielles entre les filiales internationales du Groupe et ses entités métier/support au Maroc (retail, corporate, micro-crédit, crédits conso, factoring, leasing, assistance, risques, IT, digital...). Ses interventions ont été progressivement étendues aux autres structures du Groupe à l'échelle nationale.

2. Domaines d'actions :

BCP Consulting intervient en particulier dans :

- ✓ Le renforcement de la capacité d'exécution et de transformation ;
- ✓ L'apport d'expertise pour les chantiers structurants ;
- ✓ L'assistance au pilotage des projets complexes tels que définis dans la feuille de route stratégique des filiales et le plan de développement global du Groupe ;
- ✓ La contribution à la définition des dispositifs cibles des filiales, en relation avec les marchés et fonctions centrales du Groupe ;
- ✓ Plus généralement, la réponse aux sollicitations d'accompagnement émanant des filiales.

A travers une offre de service variée et taillée sur mesure, BCP Consulting permet au Groupe de canaliser ses ambitions à l'international.

3. Quelques indicateurs clés :

Aujourd'hui, BCP Consulting c'est :

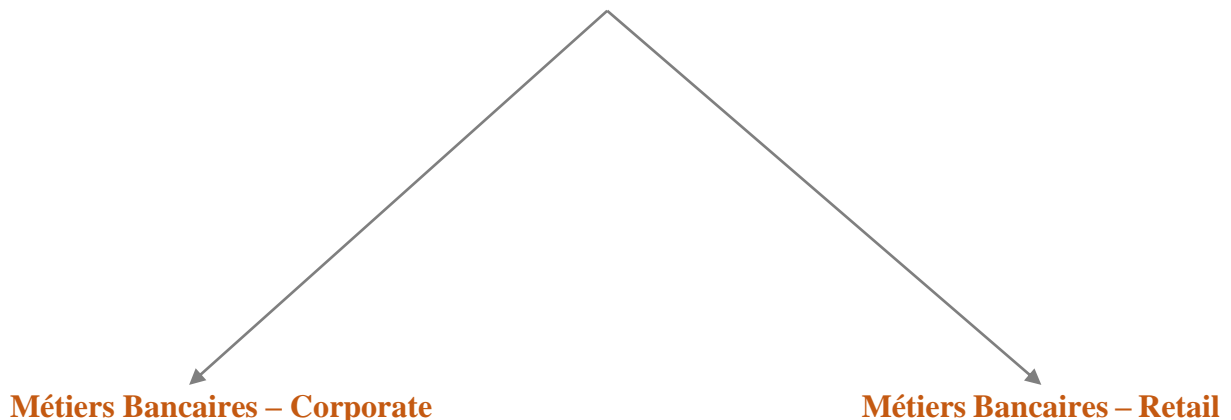
- ✚ 5 Pôles d'expertise
- ✚ 15 Collaborateurs
- ✚ 90 Projets stratégiques
- ✚ 20 Clients / Filiales bénéficiant d'un accompagnement personnalisé
- ✚ 10 Pays d'intervention



Figure 1 : Pays d'intervention de BCP Consulting

Afin de répondre aux objectifs préalablement mentionnés, BCP Consulting est organisée autour de 5 pôles d'expertise :

- ✦ **Modélisation et Ingénierie Financière** : Stabilisation de modèles financiers adaptés aux besoins et aux contextes de chaque filiale pour une évaluation appropriée des risques.
- ✦ **Conformité réglementaire et gestion des risques** : Accompagnent à la gestion efficace et durable des risques stratégiques, réglementaires, financiers, opérationnels et de conformité.
- ✦ **Système d'information et Digital** : Apport d'assistance depuis la conception stratégique à la mise en œuvre opérationnelle des solutions informatiques et digitales.
- ✦ *Excellence opérationnelle et synergies* : Accompagnement aux projets de transformation organisationnels, d'excellence opérationnelle et d'activation des synergies clés pour la refonte et/ou l'optimisation des processus.



II. Risque bancaire : Risque de crédit

Les établissements de crédit sont exposés à plusieurs risques. Certains sont spécifiques au métier du banquier et d'autres relèvent des entreprises elles-mêmes.

L'excès de risque représente la cause majeure de défaillance bancaire. Cet excès de risque est la conséquence d'une gestion et d'un contrôle inefficace de l'activité d'octroi de crédit par la banque.

La mission de l'organisme responsable de la supervision bancaire est de limiter l'ensemble des différents risques tel que le risque de crédit, de marché, et le risque opérationnel. Nous allons ici, plus mettre en avant le risque de crédit au détriment des autres, car c'est celui que nous visons à réduire à travers notre modèle.

1. Risque de crédit :

La banque par sa nature prend des risques. Il existe plusieurs types de risques auxquels elle est confrontée entre autres le risque de crédit.

Le risque de crédit est devenu une préoccupation importante des banques à la fin des années 1980, et n'a cessé d'être évalué et modélisé depuis. Il présente trois composantes : la probabilité de défaut (ou la probabilité que l'emprunteur ne respecte pas ses conditions), le taux de recouvrement au moment du défaut (loss given default) et l'exposition au risque de crédit au moment du défaut.

L'un des risques majeurs du risque de crédit est le non-remboursement qui peut constituer une source de défaillance de bancaire. Le risque de non-remboursement est essentiellement dû à une gestion et un contrôle inefficace de l'activité d'octroi de crédit. Afin d'éviter ce risque, la banque cherche des outils qui lui permettent de l'appréhender en trouvant des moyens qui assurent sa couverture. Pour cela la banque fait appel aux méthodes quantitatives qui permettent de prédire la probabilité de défaut.

2. Probabilité de défaut :

Les établissements de crédit se sont adaptés aux évolutions rapides de leurs environnements. L'instauration de techniques permettant de visualiser rapidement et efficacement les potentiels dangers sur chaque portefeuille. Avant toute chose la banque doit identifier et évaluer les risques avant de pouvoir les traiter.

Le risque de défaut est par définition la probabilité qu'une entreprise se trouve en position de défaut dans un certain horizon temporel.

La probabilité de défaut d'un client est le risque que sa note baisse pendant la période à venir. Plus sa note initiale est bonne, moins cette probabilité est importante. La PD est calculée par la banque sur des données historiques accumulées (au moins 2 ans pour que le calcul soit fiable). Elle mesure la probabilité d'occurrence d'un défaut sur une contrepartie donnée dans un horizon donné. C'est un élément qui permet de mesurer le risque lié à l'emprunteur.

3. Evénement de défaut :

La définition de l'événement de défaut est directement liée à l'estimation de la probabilité de défaut. Le défaut est défini au niveau d'un client et non d'un crédit. Lorsqu'un client rentre en défaut à cause d'un crédit, il contamine tous les autres crédits qu'il possède.

Selon la BIS, si le contrat (un ou plusieurs) d'un client atteint plus de 90 jours d'impayés et que le montant dû (sur l'ensemble des contrats ayant atteint 90 jours d'impayés) est au-delà d'un certain seuil de matérialité, alors le client est considéré en défaut. La contagion est appliquée au niveau du client (tous ses crédits entrent en défaut).

III. De Bâle I vers Bâle II :

Le comité de Bâle sur la supervision bancaire a été institué en 1975 en réponse à la croissance des échanges bancaires internationaux et donc à la nécessité de disposer de standards et principes communs pour renforcer la solidité et la stabilité du système bancaire et financier international. Le comité fait jouer un rôle central aux fonds propres et ses travaux visent à proposer des normes régissant le niveau et la gestion des fonds propres à l'intention des banques ayant une activité internationale importante.

1. L'approche classique de la gestion des risques : **RATIO COOKE** (Bâle I) :

Créé en 1974 par les gouverneurs des Banques centrales des pays du G10, le Comité de Bâle s'est donné pour mission de définir des règles visant à améliorer la stabilité du système bancaire international. Cet objectif impliquant en premier lieu de limiter le risque de faillite des banques, le Comité s'est d'abord concentré sur le risque de crédit en fixant un seuil minimal à la quantité de fonds propres des banques qui servent à couvrir les pertes subies sur les crédits accordés.

L'accord de Bâle I, qui répond à cet objectif, est un ensemble de recommandations formulées en 1988 par le Comité de Bâle, dont la principale donne est la définition du ratio Cooke. Ce ratio détermine le niveau minimum de fonds propres susceptible de couvrir les risques auxquels s'expose l'établissement bancaire. Le calcul est effectué de la sorte :

$$\text{Ratio Cooke} = \frac{\text{Fonds propres}}{\text{Risque de crédit} + \text{Risque de marché}} \geq 8\%$$

Ce dernier exige que le ratio des fonds propres réglementaires d'un établissement de crédit rapporté à l'ensemble de ses engagements de crédit ne soit pas inférieur à 8%.

Malgré cette première étape vers une réglementation plus stricte des activités bancaires, Bâle I ne couvrait que les risques de crédit et de marché, et ne proposait aucune mesure concernant les risques opérationnels.

Par conséquent et afin de palier à ces insuffisances, le comité de Bâle publiait en 2004 un nouveau cadre réglementaire dit Bâle II. En effet, l'objet essentiel de Bâle II demeure : le renforcement de la stabilité du système bancaire.

2. L'approche dynamique de la gestion des risques : RATIO MC DONOUGH (Bâle II) :

Le 26 juin 2004 étaient publiées les recommandations, dites Bâle II, mettant en place le ratio Mc Donough qui devait progressivement remplacer le ratio Cooke, le taux restait alors inchangé à 8% mais devait tenir compte des risques de crédit, marché et opérationnels. Les recommandations de Bâle 2 ont été mises en place jusqu'au 1er janvier 2008.

$$\text{Ratio Mc Donough} = \frac{\text{Fonds Propres}}{\text{Risques (Crédit + Marché + Opérationnel)}} \geq 8\%$$

La réforme Bâle II est plus complète et définit une mesure plus pertinente du risque. Ce nouveau cadre réglementaire s'appuie essentiellement sur 3 piliers afin de mieux appréhender les risques bancaires.

Pilier 1 : Exigences minimales de fonds propres	Pilier 2 : Surveillance par les autorités prudentielles	Pilier 3 : Transparence et discipline de marché
<ul style="list-style-type: none"> ◆ Risque de crédit (nouvelles approches de calcul) ◆ Risque de marché (inchangé) ◆ Risque opérationnel (nouveau) 	<ul style="list-style-type: none"> ◆ Evaluation des risques et dotation en capital spécifiques à chaque banque ◆ Communication plus soutenue et régulière avec les banques 	<p>Obligation accrue de publication (notamment de la dotation en fonds propres et des méthodes d'évaluation des risques)</p>

Tableau 1 : Les piliers de Bâle II

3. Approche Quantitative de la mise en place d'un SNI des entreprises telle que recommandée par Bâle II :

La réforme des ratios de solvabilité bancaire élaborée par le Comité de Bâle (Bâle II) vise à mettre en adéquation les fonds propres des banques avec les risques qu'elles prennent. Ces nouvelles règles plus orientées vers la notion de risque réel, permettront aux banques de recourir à leurs propres modèles de notation de leurs clients, ce qui représente la principale avancée de Bâle II.

Les modifications apportées par Bâle II au système d'évaluation ont pu être considérées comme un point majeur de la réforme. En d'autres termes, les accords de Bâle II réaffirment l'importance de la notation financière mais ils prévoient que cette notation peut être soit effectuée par des agences (Standard & Poor's, Fitch, Moody's...) très décriées aujourd'hui, soit réalisée en interne avec des méthodologies propres.

Chapitre 2 : Elaboration d'un modèle de notation interne

– Cadre théorique

Sous l'impulsion de la réglementation prudentielle Bâle II et Bâle III, les banques sont amenées à développer des systèmes internes de notation pour la mesure du risque de crédit. Ainsi, la notation de crédit occupe aujourd'hui une place sans précédent dans les pratiques de ces institutions.

I. Notation externe Vs notation interne :

Au titre du risque de crédit, la réglementation Bâle II offre aux banques le choix entre deux grandes méthodes pour le calcul de fonds propres exigibles. La première approche consiste à évaluer ce risque d'une manière standard et se distingue nettement de son prédécesseur Bâle I par la reconnaissance des évaluations externes. La seconde approche permet aux banques d'utiliser leur système interne de notation afin de saisir les caractéristiques réelles de l'emprunt et d'introduire plus de sensibilité dans la mesure du risque de crédit.

1. Notation externe :

La notation, bien qu'elle puisse être effectuée par le biais d'un modèle de risque propre à la banque, peut être réalisée par un organisme externe.

Les agences de notation (Credit Rating Agency) sont des entreprises privées dont l'activité principale consiste à évaluer la capacité des émetteurs de dette à faire face à leurs engagements financiers.

Force est de reconnaître que l'objectif dévolu par le régulateur aux évaluations externes, fournies essentiellement par les grandes agences de notation (Fitch Rating, Moody's Investors Service et Standard & Poor's Rating Services), est d'introduire plus de sensibilité aux risques encourus par les banques adoptant l'approche standard.

La notation (rating) donne une opinion sur la capacité d'un émetteur à remplir ses obligations vis-à-vis de ses créanciers, ou d'un titre à générer les paiements de capital et

d'intérêts conformément à l'échéancier prévu. Les entités notées sont donc potentiellement tous les agents financiers ou non financiers émetteurs de dette : états, organismes publics ou semi-publics, établissements financiers, entreprises non financières.

<i>Moody's</i>	<i>Standard and Poor's</i>	<i>Fitch ratings</i>	<i>Signification</i>
Aaa	AAA	AAA	Le risque est quasi nul, la qualité de la signature est la meilleure possible. La sécurité est optimale.
Aa1, Aa2, Aa3	AA+, AA, AA-	AA+, AA, AA-	Quasiment similaire à la meilleure noté, l'émetteur noté AA est très fiable.
A1, A2, A3	A+, A, A-	A+, A, A-	Bonne qualité mais le risque peut être présent dans certaines circonstances économiques.
Baa1, Baa2, Baa3	BBB+, BBB, BBB-	BBB+, BBB, BBB-	Solvabilité moyenne et la qualité est inférieure.
Ba1, Ba2, Ba3	BB+, BB, BB-	BB+, BB, BB-	A partir de cette note, l'affaire commence à être spéculative. Le risque de non remboursement est plus important sur le long terme.
B1, B2, B3	B+, B, B-	B+, B, B-	La probabilité de remboursement est incertaine. Il subsiste un risque assez fort. Cela reflète une situation hautement spéculative.
Caa	CCC	CCC	Risque très important de non remboursement sur le long terme.
Ca	CC	CC	Très proche de la faillite, emprunt très spéculatif.
C	C	C	Situation de faillite de l'emprunteur.
	D	DDD	Défaut
		DD	Défaut
		D	Défaut

Tableau 2 : Symboles de notation des agences de notation

2. Notation interne :

Les banques peuvent, si elles le désirent, fixer elles-mêmes en interne les pondérations des risques. L'idée est que les banques sont les mieux à même d'évaluer leurs risques. Il est possible alors de distinguer trois options pour le calcul du risque de crédit : L'approche standard et deux approches distinctes fondées sur les notations internes (NI) : l'approche fondation et l'approche avancée. L'approche standard, qui correspond pour l'essentiel à l'approche de Bâle I, consiste à utiliser des systèmes de notation fournis par des organismes externes d'évaluation du crédit alors que l'approche notation interne dite « NI », implique que l'établissement de crédit attribue lui-même une note ou score à son client en fonction de différents paramètres.

L'approche NI, qui repose sur des évaluations internes réalisées par la banque, donne lieu à une distinction entre l'approche des notations internes dite IRB simple (Foundation internal rating-based) et l'approche des notations internes dites IRB avancée (Advanced internal rating-based).

♦ **L'approche IRB fondation (simple) :** Cette méthode prévoit que les banques utilisent leurs évaluations internes de la probabilité de défaillance (PD) de leurs clients de façon à déterminer les exigences de fonds propres. Les autres données nécessaires au calcul du risque de crédit (pertes en cas de défaillance (LGD), exposition anticipée en cas de défaillance (EAD) et maturité (M)) seront fournies par les autorités de tutelle.

L'adoption de cette approche ne pourra se faire qu'aux conditions suivantes :

- 1 an d'utilisation des modèles de calcul des PD
- 2 ans d'historique des données relatives aux défaillances et 5 ans à terme
- Une validation par les autorités de tutelle qu'une part déterminante des encours, mais également, représentative de la diversité des métiers du groupe est traitée sous le régime de l'IRB Fondation.

♦ **L'approche IRB avancée :** Elle se distingue principalement par l'utilisation étendue par les banques de leurs propres estimations pour certaines données de base alors que celles-ci sont fournies par l'autorité de surveillance pour l'approche IRB simple. Il y avait une volonté de mieux évaluer les risques bancaires et imposer un dispositif de surveillance prudentielle et de transparence.

Cette méthode prévoit que les banques utilisent leurs évaluations internes du risque de crédit (probabilité de défaillance (PD), pertes en cas de défaillance (LGD), exposition anticipée en cas de défaillance (EAD) et maturité (M)) pour déterminer les exigences de fonds propres.

L'adoption de cette méthode est plus contraignante que l'IRB Fondation :

- 3 ans d'utilisation des modèles pour le calcul des PD, LGD, EAD et M
- 7 ans d'historique des PD, LGD, EAD et M
- Une validation par les autorités de tutelle qu'une part déterminante des encours, mais également, représentative de la diversité des métiers du groupe est traitée sous le régime de l'IRB Avancée.

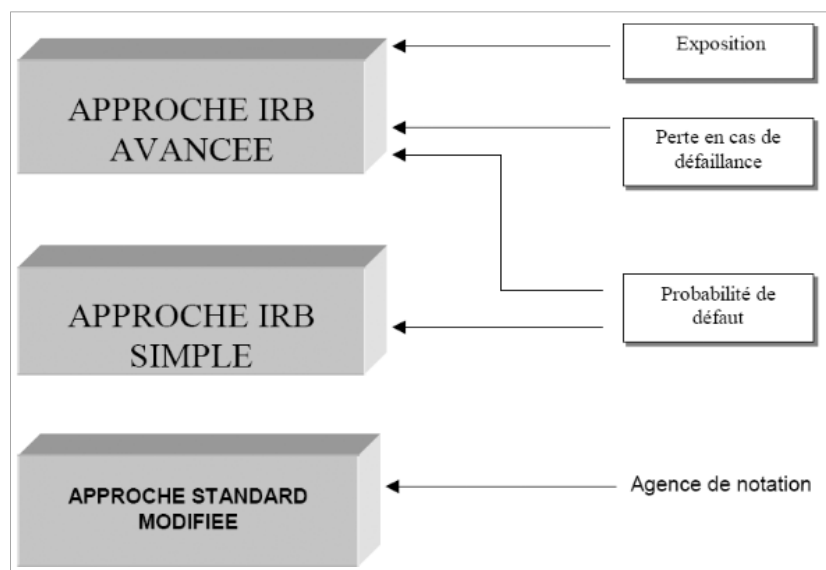


Figure 2 : Approches de notation interne et externe

II. Méthodologie d'élaboration d'un modèle de score :

La défaillance des entreprises a fait depuis plusieurs années, l'objet de nombreux travaux. Elle est aujourd'hui, particulièrement remise au goût du jour avec l'obligation pour les banques de noter leurs créances, dans le respect de la nouvelle réglementation (Bâle II). La grande majorité des travaux s'appuie sur des outils d'analyse statistique de grandeurs comptables et de ratios financiers pour discriminer les entreprises saines des entreprises défaillantes. Elle débouche sur un **calcul de score**.

Un score est un indicateur de synthèse censé donner en un chiffre, le degré de défaillance possible d'un débiteur. Il permet d'aider à la notation des créances, par une approche quantitative du risque de défaillance. La notation est entendue comme une opinion indépendante et publique sur la qualité de crédit d'une entité.

1. Architecture générale d'un système de notation interne :

Afin d'être éligible à l'approche IRB, Le comité de Bâle préconise aux banques de satisfaire les exigences minimales en termes de système de notation et d'être en mesure d'appliquer ces dispositifs de manière adaptée à chaque segment de clients (PME, grosse entreprise, particulier, etc.).

L'expression système de notation recouvre : « *l'ensemble des processus, méthodes, contrôle ainsi que les systèmes informatiques et de collecte des données qui permettent d'évaluer le risque de crédit, d'attribuer des notations internes et de quantifier les estimations de défaut et des pertes.* ». L'architecture générale d'un système interne de notation s'articule autour d'un certain nombre d'aspects méthodologiques auxquels les banques doivent porter une attention particulière. La description des différentes étapes nécessaires à la construction d'un système de notation se décline comme suit :

- **La définition des pertes :** La première étape consiste pour la banque d'être précise quant aux concepts de pertes utilisés. Dans la plupart des cas, l'évaluation du risque de crédit résulte de la combinaison de trois paramètres. Le premier étant le risque de défaut (PD) de l'emprunteur qui se matérialise par la probabilité que ce dernier soit dans l'incapacité de faire face à l'une quelconque de ses dettes sur un horizon de temps déterminé. Le deuxième peut être appréhendé comme le risque de ne se voir récupérer une fraction de la dette en cas de défaut de l'emprunteur (LGD). Le dernier paramètre n'est que l'exposition au défaut (EAD) qui représente pour chaque engagement de la banque le montant de la créance dû par l'emprunteur en défaut. Toutefois, Cette procédure de notation n'est pas exemptée de critiques. Comme le montre Edward (2012), l'une des limites capitales de ces modèles réside dans l'étroitesse de l'échantillon de base ayant servi à leur conception.

$$EL = PD \times LGD \times EAD$$

EL : Expected Loss ou Perte attendue

PD : Default Probability

LGD : Loss Given Default

EAD : Exposure At Default

- **La constitution d'une grille de notation :** A ce stade, la banque est amenée à détailler le nombre et la signification des classes de risque considérés selon une échelle de notation qui devrait être hiérarchisée en fonction de la qualité de crédit. L'élaboration d'une grille de notes requerrait préalablement la scission du portefeuille de créances en deux principales catégories, en l'occurrence, les emprunteurs sains et ceux en défaut. La première catégorie s'attèle aux créances dont le règlement s'effectue à l'échéance et qui sont détenues par des emprunteurs dont la capacité à honorer leurs engagements, immédiats et/ou futurs, ne présente pas de motif d'inquiétude. La seconde composante désigne les créances qui présentent un risque de non recouvrement total ou partiel eu égard à la détérioration de la capacité de remboursement immédiate et/ou future de l'emprunteur.
- **Le processus de notation :** L'étape cruciale dans cette architecture reste le processus opérationnel « operational design » de la notation interne. Les aspects fondamentaux de cette procédure résident, avant tout, dans le rôle important de la direction responsable de proposer la première appréciation du rating. Il s'agit notamment des centres d'affaires dont les responsables des dossiers de crédit se penchent sur une analyse profonde de la contrepartie au travers d'une sélection de questions et documents collectés, et appuyés dans leur réflexion par des outils d'aide à la décision de notation (des analyses statistiques comme le scoring financier).
- **La validation du dispositif de notation :** Au-delà des aspects précités, la robustesse d'un système de notation conformément aux exigences de la réglementation bâloise devrait également faire l'objet d'un contrôle de validation avant toute généralisation opérationnelle. A cet effet, des techniques de simulation de type « backtesting » sont

mises en œuvre afin de tester la fiabilité et la capacité prédictive du dispositif et ce, par une analyse comparative des anticipations reflétées dans la note finale avec les données observées ex-post d'un échantillon de clients.

2. Le modèle de crédit scoring :

Pour une banque, la gestion du risque que représente le crédit est un aspect fondamental de leur activité. On ne prête pas à tout le monde, il faut des garanties de la part des demandeurs de crédit. Le problème, c'est que bien souvent ces garanties présentées par les demandeurs ne sont pas suffisantes, la banque a besoin de plus de données pour pouvoir se décider à prêter de l'argent, d'où le besoin de faire un scoring.

Le crédit scoring, ou encore scoring d'octroi, est un des outils mis en œuvre lors de l'analyse risque d'une demande de crédit par les prêteurs. Méthode statistique adaptée à une pratique massive du crédit, son impartialité est souvent citée parmi ses vertus par l'industrie. Elle génère toutefois des refus de crédit qui n'auraient pas lieu d'être : mise en lumière d'une limite méthodologique.

2.1. Principes généraux du crédit scoring :

La performance et la robustesse des modèles de crédit scoring dans la classification des emprunteurs, reste, à ce jour, une question ouverte. Cette performance dépend essentiellement des procédures suivies lors de la construction des modèles de notation en question et du degré de connaissance de ces utilisateurs, une fois le modèle mis en place :

- Le modèle doit contenir un maximum d'informations ;
- Selon le comité de Bâle, les données historiques qui couvrent une période assez longue, doivent couvrir un cycle économique ;
- Les coefficients de la fonction score doivent être significatifs et conformes à la logique comptable ;
- L'échantillon de construction sur lequel est estimé, le modèle doit être homogène ;

- L'échantillon de construction doit comprendre un nombre assez grand d'emprunteurs (emprunteurs en défaut ou non) pour qu'il soit représentatif du portefeuille du crédit ou d'un segment de portefeuilles ;
- Pour faire face à la dérive temporelle, il sera nécessaire d'examiner en détail la situation financière du client ;
- La performance du modèle est jugée en fonction des taux de bon classement, c'est la raison pour laquelle le modèle doit prévoir le défaut ;
- Les performances du modèle doivent être stables à un instant donné (en réalisant des tests sur des populations différentes) et au cours du temps (la prévision reste valable à un horizon compris entre 18 et 24 mois), au-delà, il faut estimer un nouveau modèle car il est exposé au changement de la population des emprunteurs sains ainsi que leurs caractéristiques.

2.2. Les types de modèles de crédit scoring :

Avant d'aborder la méthodologie d'élaboration d'un modèle de score, il est important de mettre en évidence les différents types de modèles existants, car ceux-ci détermineront le cheminement à suivre pour développer le modèle. Il y a deux types de modèles, en fonction de l'obtention des scores qui sont : les modèles déductifs et les modèles empiriques.

2.2.1. Les modèles déductifs ou à priori :

Un système de crédit scoring déductif consiste à identifier des *variables prédéterminées* représentant des critères de risques (des effets de diverses caractéristiques d'un emprunteur), et de leur attribuer des points (ou des poids). Ces points sont déterminés par des experts de crédit ou des décideurs en fonction de leurs expériences professionnelles dans le domaine. La somme de ces points constituera un « score ». Cette approche n'utilise aucune méthode statistique, ce qui permet de les qualifier de « subjectif » car ils se basent sur des avis d'experts, et doivent, à ce titre, être donc utilisés avec prudence.

2.2.2. Les modèles empiriques ou basés sur l'historique :

Les systèmes empiriques de crédit scoring reposent sur une analyse statistique et objective des critères de risque. Où les méthodes statistiques s'appuient principalement sur des techniques de classification. Ils Permettent une évaluation approfondie de l'emprunteur. La sélection des variables discriminantes et la détermination de la fonction score (et donc le calcul du score), se réfèrent à des données similaires aux crédits déjà octroyés (une base de données historiques) selon le principe « le passé est la meilleure estimation du futur ». L'utilité de ce type de modèle se reflète dans sa prise en considération de plusieurs critères simultanément, sans aucun jugement subjectif, et de l'interdépendance existant entre ces mêmes critères.

2.3. Démarche de construction d'un modèle de crédit scoring :

Le développement d'un modèle de scoring dépend du type de ce dernier. Ainsi, dans le cadre des modèles empiriques, la conception d'un modèle de credit scoring suit une procédure relativement standard. Elle est fondée sur l'observation ultérieure de l'avenir des entreprises. Il convient de classer les emprunteurs en deux populations distinctes, l'une regroupant des emprunteurs en défaut, et l'autre des emprunteurs n'ayant pas fait défaut. Et donc, la première étape de cette conception consiste à choisir un critère de défaut, par un préjugement de la défaillance des entreprises de l'échantillon. L'objectif est de construire un modèle statistique qui crée une relation dichotomique entre les variables les plus discriminantes et le fait d'avoir connu la défaillance ou non, après avoir sélectionné ces variables individuellement et vérifié leurs degrés de signification. Ce qui permet de déterminer à un instant donné du temps la probabilité de défaut. « Si le modèle ne fournit pas directement une probabilité de défaut, il peut être nécessaire de transformer le score (qui exprime le risque de défaillance) formellement en probabilité d'occurrence. D'abord, la procédure est d'affecter le score en classe de risque selon le théorème de Bayes (discrétisation). Ensuite un traitement statistique est effectué ». Par conséquent, des étapes fondamentales sont à la base de la mise en œuvre d'un modèle de crédit scoring :

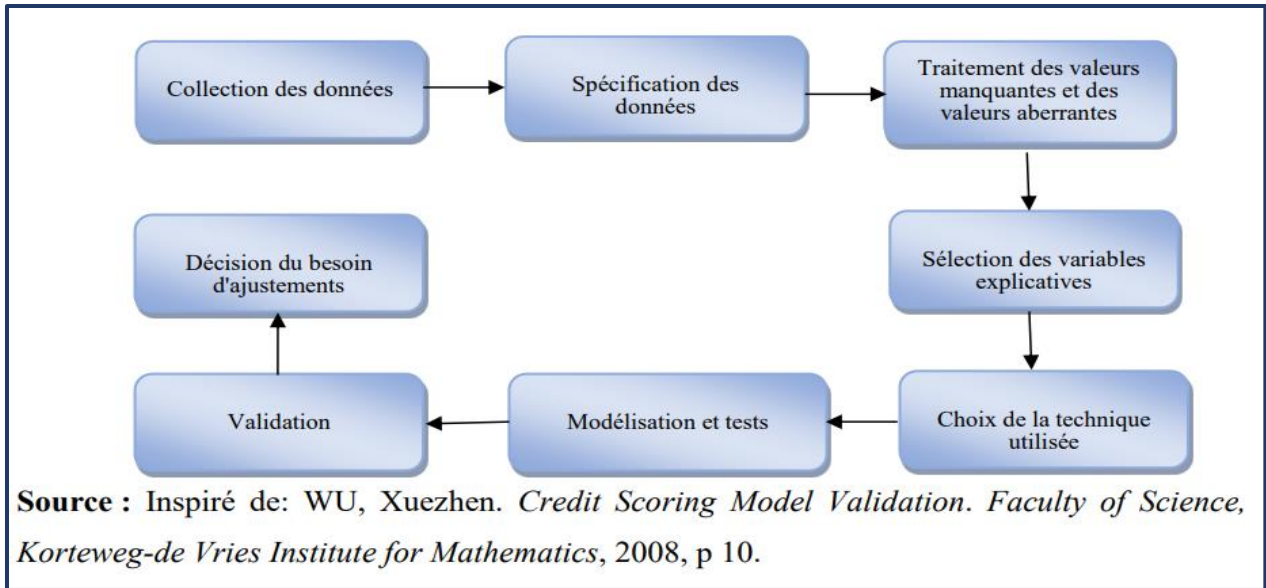


Figure 3 : Etapes de la conception d'un modèle de credit scoring

2.3.1. Collection des données :

Cette étape comporte le choix du critère de défaut et de la population à analyser :

♦ **Choix du critère de défaut :** La première étape d'un projet de développement d'un modèle de score est de définir l'événement du défaut. « En analyse du risque de crédit, cet événement peut être de deux natures. Il peut s'agir de la faillite, qui est un événement objectif de caractère juridique. Ou du défaut, qui est le non-respect d'un engagement de crédit ». Un défaut peut prendre plusieurs formes, et son appréciation comporte une part de subjectivité. Par conséquent, dans son Accord sur les fonds propres, le Comité de Bâle a donné une définition de référence de l'événement du défaut et a annoncé que les banques devraient utiliser cette définition réglementaire pour estimer leurs systèmes de notation internes. Selon cette définition, un défaut intervient lorsque l'un des événements suivants survient :

- Lorsqu'un débiteur est dans l'incapacité de rembourser, la banque estime qu'il est peu probable que le débiteur paie intégralement ses obligations de crédit au groupe bancaire ;
- Report du paiement associé à un événement de type abandon de créances, provision spécifique ou restructuration en période de difficultés ;
- Un retard de paiement de plus de 90 jours ;
- L'emprunteur est juridiquement en faillite.

♦ **Construction des échantillons (Caractéristiques d'entrée) :** Après avoir déterminé le critère de défaut, l'étape suivante consiste à présélectionner des données historiques sur ces défauts (caractéristiques d'entrée à inclure dans l'échantillon). Ces caractéristiques doivent décrire les facteurs de risque de crédit les plus importants, à savoir l'effet de levier, l'utilisation des actifs, la liquidité, la rentabilité et la performance opérationnelle ; et de constituer un échantillon composé d'un nombre suffisant d'emprunteurs en défaut (défaillants), et autre d'emprunteurs non défaillants (sains). Ces échantillons doivent être doublement représentatifs (représentatifs de la relation macroéconomique et la relation entre entreprises défaillantes et non défaillantes) de la population totale à laquelle le modèle est censé être appliqué. Il est également nécessaire qu'ils regroupent des emprunteurs appartenant à des populations homogènes (Ayant des caractéristiques comparables) pour ne pas être affectés par des différences structurelles. Ceci conduit, le plus souvent, à construire des modèles de scores spécifiques pour des secteurs particuliers (modèle de score par industrie ou par filière).

2.3.2. Spécification des données :

♦ **Définir l'horizon temporel :** L'horizon temporel fait référence à la période d'estimation de la probabilité de défaut, c'est une période historique (plus ou moins longue) avant la faillite. Le choix de cet horizon est un compromis entre la fonction assignée au modèle élaboré et la disponibilité des données traitées. Il présente une décision clé pour la construction d'un modèle de crédit scoring, car il varie en fonction de l'objectif de développement du modèle (estimation de la probabilité de défaut à court, à moyen ou à long terme). Pour la plupart des banques, il est courant de choisir une année (en utilisant l'information de l'année précédente N-1 pour prévoir les défaillances de l'année en cours N) comme horizon de modélisation, ce qui est suffisamment long pour que les banques prennent des mesures pour atténuer le risque de crédit ; et d'autre part, de nouvelles informations concernant les débiteurs et les données de défaut peuvent être révélées en une année. Cependant, un horizon temporel plus long pourrait également être intéressant, en particulier lors de la prise de décisions sur l'attribution de nouveaux prêts, mais généralement, il peut y avoir un manque de données. Pour cette raison, les modèles conservent l'horizon qui prend en compte le temps requis pour obtenir l'information financière utilisée.

◆ **Répartition des données :** Puisque la construction du modèle et sa validation nécessitent des échantillons et que l'évaluation statistique de la performance d'un modèle de score prédictif peut être très sensible à l'ensemble de données, ce dernier doit être suffisamment grand pour le divisé au hasard en deux sous-ensembles, l'un pour le développement et l'autre pour la validation. Pour éviter l'inclusion de la dépendance aux données indésirables, un certain type de test hors échantillon, hors du temps et hors de l'univers doit être utilisé dans le processus de validation. Normalement, 60% à 80% de l'échantillon total est utilisé pour estimer le modèle ; l'échantillon restant de 20% à 40% est mis de côté pour valider le modèle.

◆ **Exploration des données :** Avant d'initier la modélisation proprement dite, il est très utile de calculer des statistiques simples pour chaque caractéristique, telles que la moyenne, la médiane, l'écart-type et la plage de valeurs. L'interprétation des données doit également être vérifiée. Par exemple, il faut s'assurer que « 0 » représente zéro, et non des valeurs manquantes, et confirmer que toutes les valeurs spéciales sont documentées. Cette étape vérifie que tous les aspects des données sont bien compris et offre un bon aperçu de l'entreprise.

2.3.3. Traitement des valeurs manquantes et des valeurs aberrantes :

La plupart des données de l'industrie financière contiennent des valeurs manquantes ou des valeurs aberrantes qui doivent être correctement gérées. De nombreuses méthodes permettent le traitement des valeurs manquantes, telles que la suppression de toutes les données avec des valeurs manquantes ou l'exclusion des caractéristiques des enregistrements contenant des valeurs manquantes significatives du modèle, mais cela peut entraîner une perte de données importante. Une autre méthode directe consiste à remplacer les valeurs manquantes par les valeurs moyennes ou médianes correspondantes sur toutes les observations pour la période correspondante. Bien que ces trois méthodes supposent qu'aucune information supplémentaire ne puisse être recueillie à partir de l'analyse des données manquantes, cela n'est pas nécessairement vrai et les valeurs manquantes sont généralement non aléatoires. Les valeurs manquantes peuvent faire partie d'une tendance, être liées à d'autres caractéristiques ou indiquer une mauvaise performance. Par conséquent, elles doivent être analysées en premier lieu, et si elles s'avèrent être aléatoires et performantes, elles peuvent être exclues ou imputées en utilisant

des techniques statistiques. Autrement, si les valeurs manquantes sont corrélées à la performance du portefeuille, il est préférable de les inclure dans l'analyse.

Concernant les valeurs aberrantes, qui sont des valeurs éloignées des autres pour une propriété particulière. Elles peuvent avoir un effet négatif sur les résultats de la régression. Bien que la solution la plus simple consiste à supprimer toutes les données extrêmes qui se situent en dehors de la plage normale, par exemple à une distance de plus de deux ou trois fois l'écart-type, en utilisant cette solution, il est très facile d'éliminer par erreur les entreprises défailtantes, considérées comme aberrantes. Une autre méthode appelée « winsorisation », consiste à définir des valeurs aberrantes sur un pourcentage spécifique de données.

2.3.4. Sélection des variables explicatives :

Dans cette étape de la sélection des variables, le pouvoir prédictif de chacune d'elles, sera évalué individuellement (analyse univariée). Cette sélection est délicate, il s'agit principalement des données (quantitatives et/ou qualitatives) utilisées pour la classification et la séparation entre les deux échantillons, et qui peuvent être traités par le modèle. A l'origine, un grand nombre de variables est utilisée pour construire un modèle de score. Ces variables doivent déterminer la dimension de variation de risque de défaut (la solidité financière, endettement, la rentabilité, l'évolution des délais, la gestion du cycle d'exploitation rentabilité, etc.). Parmi celles-ci, seul un petit nombre sera finalement pertinent et retenu dans le modèle (généralement, moins d'une dizaine) en fonction de leur capacité discriminante individuelle. Le choix de variables explicatives se fait par rapport aux différents types de données :

- Les informations comptables et financières, sous forme de ratios financiers offerts par l'analyse financière, constituant une série d'indicateurs du risque d'une entreprise, et de variables issues de tableau prévisionnels de flux de trésorerie ;

- Les données bancaires, qui peuvent être obtenues, en interne identifiées par la régularité du comportement de paiement des emprunteurs, ainsi que la situation de leurs soldes. Ou auprès de sources externes comme des fichiers partagés par la profession bancaire ;

- Les notations externes des emprunteurs fournies par les agences de notation financière, qui peuvent servir de benchmarks pour l'évaluation des systèmes internes de notation ;

- Les informations qualitatives, portent sur les différentes variables concernant les entreprises (position concurrentielle, options stratégiques, qualité de gestion et d'organisation, dépendance relative aux différents types de risques, ...), ou concernant les particuliers (âge, ancienneté, localisation géographique, profession, ... etc.).

Par conséquent, afin d'inclure ces données dans le modèle de score, la distribution des valeurs des variables est examinée et des transformations appropriées sont effectuées. Les variables les plus fortes sont regroupées et les ratios faibles ou illogiques seront éliminés. Cependant, la corrélation entre les variables retenues devrait également être testée, et c'est l'une des exigences de construction d'une fonction score. Parce que, si certains indicateurs fortement corrélés sont inclus dans le modèle, les coefficients estimés seront biaisés de manière significative et systématique. *En effet, les variables liées apportent en réalité la même information et sont redondantes*, il est donc possible d'éliminer certaines variables et de choisir une ou plusieurs variables, qui peuvent représenter toutes les informations contenues dans d'autres caractéristiques, basées sur des considérations à la fois statistiques et commerciales ou opérationnelles.

2.3.5. Choix de la méthode statistique :

Cette étape consiste à élaborer une règle de décision d'affectation de meilleure performance (réduction des erreurs de classement) qui soit la plus efficace possible, sur la base des échantillons et des caractéristiques des variables retenues. Traitées dans la section suivante, plusieurs méthodes (techniques) permettent la construction d'un modèle de score, y compris l'analyse discriminante de Fisher linéaire ou quadratique qui est une technique de classification issue de l'analyse des données, les modèles économétriques paramétriques comme les modèles *probit* et *logit* et les modèles de *régression linéaire*. Nous trouverons aussi des techniques d'intelligence artificielle, tel que *les réseaux de neurones*. Et des méthodes non paramétriques d'enveloppement de données, qui sont encore expérimentales mais donnent de très bons résultats en termes de classification, comme *l'Arbre de décision* (CART : classification and régression Tree), la méthode du K plus proche voisin (Knn: K-nearest Neighbor), Forêts aléatoires, la méthode du noyau, DISQUAL 2, Machines à vecteurs de support à moindres carrés (Least Square Supports Vectors Machines (LS-SVM))...etc.

Les techniques économétriques de scoring les plus répandues sont l'analyse discriminante linéaire et la régression logistique, pour leur simplicité et leur grande robustesse. Au cours de ces dernières années, le modèle Logit s'est progressivement imposé comme la méthodologie dominante, car il permet de générer directement des scores à une vitesse de mise en œuvre et des coûts de plus en plus satisfaisants.

2.3.6. Modélisation et tests :

Il reste encore, à ce stade, un grand nombre de caractéristiques potentielles à inclure dans le modèle, qui ont tendance à le surestimer. Il s'agit de la phase de construction proprement dite du modèle (construction de la fonction score) par la combinaison des caractéristiques financières (variables explicatives) utilisées dans l'analyse, et de son application en test, basée sur des procédures de tests statistiques de robustesse. Cette fonction est construite grâce aux différentes techniques de classification utilisées dans le crédit scoring, et elle permet la distinction entre les clients (bons ou mauvais). En général, l'efficacité est appréciée par *le critère du taux de bons classements*. Cette étape consiste à estimer le modèle sur des échantillons de contrôle (des échantillons test), composé d'entreprises (défaillantes et non défaillantes) différentes de celles des échantillons traités, plus la taille des populations est importante, plus la qualité de score tend à être élevée.

2.3.7. La validation :

En dernière étape, le modèle de scoring établi, il doit être validé par les méthodes classiques de l'inférence statistique. Le but de la validation est de confirmer que le modèle développé est applicable et de s'assurer qu'il n'a pas été surévalué. Fréquemment, les modèles de crédit scoring seront validés sur une période de croissance, avec l'objectif de confirmer la robustesse et la qualité du modèle. Cette étape doit passer aussi par la vérification et la conformité des coefficients du modèle de score aux principes de l'analyse financière (une augmentation d'un ratio de rentabilité doit réduire la probabilité de défaut...etc.). Deux contrôles classiques sont exécutés. Tout d'abord, il faut s'assurer que *le score est plus significatif de risque lors de l'approche de l'événement de défaut*. D'autre part, *le score doit être discriminant quelle que soit la taille de l'entreprise*. En plus des indicateurs de validation

classiques tels que le taux de bon classement, d'autres outils sont recommandés par certains statisticiens, tels que, les mesures d'entropie et les courbes et ratios de performance. Ces outils, permettent de valider un modèle de score individuellement et de comparer deux modèles afin de choisir le plus pertinent.

2.3.8. Décision du besoin d'ajustement :

Le but de la validation ultérieure est de confirmer que le modèle est toujours en vigueur au fil du temps, car des changements importants peuvent avoir eu lieu et doivent être identifiés, comme un changement dans la conjoncture économique générale ou l'état de l'entreprise. Dès sa mise en œuvre, le modèle de crédit scoring est maintenu. Plusieurs méthodes de test de performance doivent être appliquées, basées à la fois sur la capacité de classification et de prédiction, permettant de suivre sa performance. Et, si nécessaire, apporter des corrections en effectuant la même procédure que lors de sa construction sur un nouvel échantillon. Nous pouvons également nous attendre à une obsolescence naturelle des modèles. En pratique, la construction d'une fonction score repose principalement sur un processus de trois étapes. La première consiste à définir l'évènement du défaut et à constituer la population initiale (les deux échantillons d'entreprises), dans cette étape, une analyse préliminaire des données de l'échantillon sélectionné est effectuée. La seconde étape consiste à sélectionner les variables discriminantes et construire le modèle. La troisième et dernière étape est réservée à la réalisation d'une analyse statistique, dont l'objet est de tester la validité du modèle établi sur des exemples (des échantillons test). La figure suivante résume ce processus en trois phases importantes :

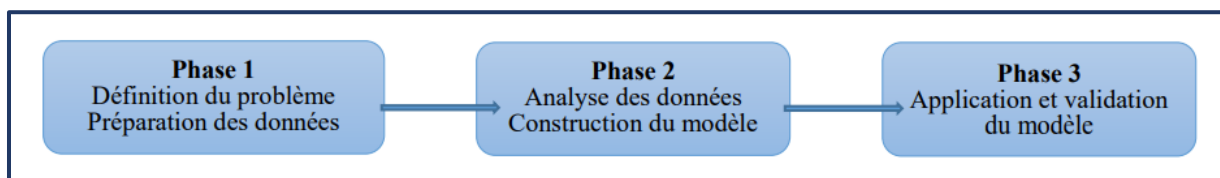


Figure 4 : Les 3 phases du processus de crédit scoring

III. Techniques de classification et de validation du modèle de scoring :

Comme nous l'avons vu précédemment, plusieurs techniques permettent l'élaboration d'un modèle de scoring, certaines sont basées sur un raisonnement probabiliste, d'autres sur la reconstitution du raisonnement humain, dans le but de calibrer une probabilité de défaillance. Nous avons choisi de concentrer notre travail sur une méthode de classification paramétrique basée sur le raisonnement probabiliste, la plus connue et utilisée dans le domaine du crédit scoring, la régression logistique et une méthode non paramétrique dite de Random Forest.

1. La régression logistique :

La régression logistique est très répandue pour les problèmes de prédiction ou d'explication d'une variable dépendante binaire (défaillance oui/non ou 1/0) à partir d'une série de variables explicatives continues, binaires ou binarisées (dummy variables). On parle dans ce cas de régression logistique binaire.

Soit un échantillon de n observations indépendantes : avec Y la variable à prédire (variable expliquée) et $X = (X_1, X_2, \dots, X_J)$ les variables prédictives (variables explicatives). Puisque nous sommes dans le cadre de la régression logistique binaire, Y représente la valeur de la variable dépendante dichotomique prenant soit la valeur zéro pour présenter l'absence, l'échec ou le « non », soit la valeur un pour présenter contrairement la présence, le succès ou bien le « oui », tandis que X représente les valeurs des différents attributs prédictifs relatifs à chaque échantillon ou participant pouvant avoir des valeurs discrètes ou continues. Effectivement, la variable Y prend deux modalités possibles $\{1,0\}$. Les variables X_j sont exclusivement continues ou binaires.

Soit $\Pi(x)$ la probabilité conditionnelle d'avoir $Y=1$ sachant que $X=x$, notée :

$$\Pi(x) = P(Y=1 | X=x)$$

S'il y a plusieurs variables prévisionnelles, l'équation de régression logistique est représentée comme suit :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_j X_j + \varepsilon, \quad j = 1, \dots, n \quad (\text{Formule standard}).$$

Vu la nature sinusoidale de la fonction logistique, l'analyse de régression logistique doit forcément transposer cette équation linéaire en expression logarithmique. En d'autres termes, au lieu de prédire un score Y, la régression logistique prédit la probabilité d'obtenir une certaine valeur cible (1 ou 0) sur Y. Deux formules alternatives et parfaitement équivalentes permettent de calculer cette probabilité :

$$\left\{ \begin{array}{l} P(Y) = \Pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \\ P(Y) = \Pi(x) = \frac{1}{1 + e^{-g(x)}} \end{array} \right.$$

Dans ces deux équations, $g(x)$ correspond à l'équation de régression linéaire conventionnelle, avec :

$$g(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_j X_j$$

Ainsi, le modèle logistique s'apparente au modèle linéaire habituellement représenté. Son expression sera donc sous la forme suivante :

$$\Pi(x) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n}}$$

$\Pi(x) \in [0,1]$ car il s'agit d'une probabilité et encore, mathématiquement parlant, le dénominateur est supérieur au numérateur :

$$1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n} > e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n}$$

Ce modèle peut être utilisé pour décrire la nature de relation entre la probabilité espérée d'un succès pour la variable réponse ($Y=1$) et les variables explicatives X , comme il peut prédire la probabilité espérée d'un succès étant donné les valeurs des variables X .

1.1. Les hypothèses de la régression logistique :

La régression logistique repose sur l'hypothèse fondamentale notée ci-dessous. Dans ce cas, nous avons pu utiliser la fonction de logarithme népérien puisqu'il s'agit d'une probabilité logistique et encore $\ln(x) = \ln \in [0,1]$.

$$\ln \frac{P(Y=1 | X=x)}{P(Y=0 | X=x)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_j X_j$$

Autrement :

$$\ln \frac{\Pi(x)}{1 - \Pi(x)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_j X_j$$

L'expression mentionnée ci-dessus est appelée Logit. Cette dernière prouve qu'il s'agit bien d'une régression logistique. En effet, la loi de probabilité est spécifiée à partir d'une « loi logistique ». Elle prouve d'autre part « la régression » car son but principal est de montrer une relation de dépendance entre une variable à expliquer et une série de variables explicatives.

Hypothèses de la régression logistique :

- La régression logistique suppose qu'il y a une multi-colinéarité minimale ou nulle parmi les variables indépendantes.
- La régression logistique suppose que les variables indépendantes sont linéairement liées au log des cotes (logit).
- La régression logistique nécessite généralement un échantillon de grande taille pour prédire correctement.

- La régression logistique qui a deux classes suppose que la variable dépendante est binaire et la régression logistique ordonnée nécessite que la variable dépendante soit ordonnée.
- La régression logistique suppose que les observations sont indépendantes les unes des autres.

1.2. Méthodes de sélection des variables :

La sélection de variables est un processus qui permet de « sélectionner » un sous-ensemble de variables considérées par le processus comme pertinentes. Les données d'entrée du processus sont constituées par l'ensemble initial de variables qui forment l'espace de représentation et l'ensemble des données d'apprentissage du problème étudié.

Le processus de sélection de variables se décompose de la manière suivante :

- A partir de l'ensemble initial des variables, le processus de sélection détermine un sous-ensemble de variables qu'il considère comme les plus pertinentes ;
- Le sous-ensemble est ensuite soumis à une procédure d'évaluation. Cette dernière permet d'évaluer les performances et la pertinence du sous-ensemble ;
- En fonction du résultat de la procédure d'évaluation, un critère d'arrêt du processus détermine si le sous-ensemble de variables peut être soumis à la phase d'apprentissage. Si tel est le cas, le processus de sélection s'arrête, sinon, un autre sous-ensemble de variables est généré. Les principaux enjeux et conséquences de la sélection de variables sont divers :
 - La sélection de variables va dans un premier temps nous permettre de déterminer les variables considérées comme pertinentes ;
 - La sélection de variables nous permet de supprimer le bruit généré par certaines variables ;
 - Les variables redondantes sont également supprimées ;
 - La taille de l'espace de représentation est ainsi réduit. Le coût de calcul de la phase d'apprentissage est également réduit.

1.2.1. La forward Selection (FS) :

Cette stratégie part d'un ensemble vide. Les variables sont ajoutées une à une. A chaque itération, la variable optimale suivant un certain critère est ajoutée. Le processus s'arrête soit quand il n'y a plus de variable à ajouter, soit quand un certain critère est satisfait. Une fois qu'une variable a été ajoutée, la FS ne peut la retirer.

1.2.2. La Backward Elimination (BE) :

Cette stratégie part de l'ensemble initial de variables. A chaque itération, une variable est enlevée de l'ensemble. Cette variable est telle que sa suppression donne le meilleur sous-ensemble selon un critère particulier. Une fois la variable supprimée, il est impossible de la réintégrer.

1.2.3. La Méthode Mixte (STEPWISE) :

Il est également possible d'utiliser une variation de l'ordre partiel des variables : Devijver et Kittler définissent un opérateur qui ajoute k ($k < p$) variables et en enlève une. La première décision à prendre est donc le point de départ de la recherche, il peut être de trois sortes :

- Un ensemble vide : il s'agit de la Forward Stepwise Selection ;
- Un ensemble complet : Backward Stepwise Elimination ;
- Un ensemble d'attributs choisis aléatoirement.

Ces méthodes permettent de pallier au problème de l'irrévocabilité de la suppression ou de l'ajout d'une variable, problème présent dans les deux autres directions de recherche. En effet, l'importance d'une variable peut se modifier ultérieurement. Ces méthodes autorisent l'ajout et la suppression d'une variable de l'ensemble des variables à n'importe quelle étape de la recherche autre que la première ou la dernière.

1.3. Validation et évaluation du modèle :

L'objectif escompté de cette étude est d'avoir le modèle le plus parcimonieux avec la plus grande performance prédictive possible. À cet effet, nous ne retiendrons que les variables explicatives significatives, dont les explications respectives du défaut, prendront un sens

économique. Ainsi, divers tests et statistiques seront mis à contribution pour valider notre spécification finale et évaluer sa performance. Cette section répertorie les mesures, tests et processus qui seront utilisés pour cette étape importante du projet.

1.3.1. Validation :

La validation d'un modèle est indissociable de tout processus d'estimation. C'est une étape cruciale dans l'élaboration de tout modèle de prédiction. À cet effet, diverses procédures et métriques qui permettent de valider un modèle sont utilisées dans la littérature.

1.3.1.1. Test du rapport de vraisemblance :

C'est un test de spécification du modèle qui compare généralement deux modèles. Basé sur les fonctions de log vraisemblance dans STATA, la statistique test qui est associée correspond à la différence entre les logs de vraisemblance respectifs du modèle dit contraint et du modèle sans restriction. La technique d'estimation par maximum de vraisemblance utilisée pour dériver les paramètres estimés de nos modèles maximise la fonction de vraisemblance. L'idée sous-jacente à ce test est que l'omission de toutes variables explicatives pertinentes (modèle contraint) impliquera une diminution importante de la log vraisemblance. Ainsi, ce ratio devrait traduire dans quelle mesure l'omission de certaines variables viendra affecter notre inférence indiquant à cet effet, la pertinence d'une spécification par rapport à une autre.

La statistique test correspondante, du nom de ratio de vraisemblance est la suivante :

$$LR = 2(\log likelihood_{sans\ restriction} - \log likelihood_{avec\ restriction})$$

Cette statistique est comparée à une chi-carré dont le degré de liberté est donné par le nombre de restrictions, dans notre cas au nombre de variables explicatives omises. La règle de décision est que si cette statistique est plus grande que le quantile associé au chi-carré, on rejette l'hypothèse nulle qui correspond au modèle restreint. De façon équivalente, plus le log vraisemblance est petit, meilleure est l'estimation et meilleur est le modèle estimé. L'avantage

de ce test est qu'il assure la plus grande puissance selon le théorème de Neyman-Pearson.

1.3.1.2. Test de stabilité :

Ce test vise à déterminer dans quelle mesure les valeurs aberrantes affectent la qualité de notre inférence. Le principe dans le présent travail est d'éliminer les valeurs aberrantes de l'échantillon et de voir par la suite, si les variables explicatives significatives de notre modèle final le demeurent. Pour ce faire, Simard a choisi d'éliminer les seules valeurs aberrantes de la variable choisie comme proxy de la taille des entreprises dans l'échantillon, soit les ventes. Ces valeurs atypiques seront éliminées au seuil de 5% et 1% respectivement. Ces spécifications nous permettront de capter la dynamique de stabilité du modèle dépendamment du niveau de conservatisme dans le traitement des valeurs atypiques.

1.3.1.3. Mesure AIC et BIC :

Toujours dans une perspective de validation de notre spécification finale, certaines métriques seront mises à contribution.

La première mesure de validation que nous allons aborder est la mesure AIC du nom de « Akaike Information Criteria ». Le critère d'information d'Akaike, de son nom français, mesure la qualité du modèle. Plus précisément, elle permet de pénaliser les modèles en fonction du nombre de paramètres choisis afin de satisfaire le critère parcimonieux, l'idée étant de choisir le modèle avec le plus bas critère. Cette analyse va parfaitement dans le sens de l'exercice de validation de notre modèle final.

La mesure AIC est définie comme suit :

$$AIC = 2K - 2 \log L$$

Où :

L = vraisemblance du modèle estimé

K = nombre de paramètres log correspond au logarithme népérien

La deuxième mesure de validation est la mesure BIC, du nom de « Bayésien information criteria ». Cette mesure a été dérivée de la mesure AIC et tient compte, en plus du nombre de variables, de la taille de l'échantillon également. Elle pénalise à cet effet plus que la mesure AIC pour l'ajout de nouvelles variables.

La mesure BIC est définie comme suit :

$$BIC = K \log N - 2 \log L$$

Où :

L = vraisemblance du modèle estimé

K = nombre de paramètres

N = taille de l'échantillon log correspond au logarithme népérien

La règle de décision est la même que pour le critère AIC, choisir le modèle avec la mesure BIC la plus basse.

Ces deux mesures s'inscrivent parfaitement dans l'optique de dériver le modèle le plus parcimonieux avec la plus grande précision. En effet, on sait que la précision d'un modèle augmente avec le nombre de variables explicatives pertinentes. Or, ces deux mesures pénalisent pour l'ajout de nouvelles variables de sorte qu'elles établissent le seuil auquel, aucune précision supplémentaire n'est apportée au modèle avec l'ajout de variables non pertinentes.

1.3.1.4. Test de significativité :

La spécification finale ne tenant compte que des variables significatives, nous allons utiliser le test de Wald pour vérifier la significativité des coefficients de ces variables. Cet exercice vient en support à l'objectif de ne retenir que les variables qui expliquent le défaut dans notre modèle de prédiction. Ce test s'aligne également avec celui du ratio de vraisemblance.

La statistique test de **Wald** suit une distribution Khi-2 sous l'hypothèse que le coefficient estimé est nul. Cette statistique test est obtenue en comparant la valeur du maximum de vraisemblance du coefficient de la variable testée avec l'écart-type du dit coefficient.

Les hypothèses correspondantes sont formulées comme suit :

Hypothèse nulle : $\beta_1 = \beta_2 = \dots = \beta_n = 0$

1.3.2. Evaluation :

L'évaluation du modèle permet de mesurer la capacité d'une variable ou d'un ensemble de variables exogènes à distinguer les classes de la variable endogène. L'optimalité d'un sous-ensemble est relative à la fonction d'évaluation utilisée. Dash et Liu considèrent que ces fonctions d'évaluation ou critères peuvent être regroupées en cinq catégories qui sont les suivantes :

- Les critères d'information : C'est la quantité d'information apportée par une variable sur la variable endogène. La variable, ayant le gain d'information le plus élevé, sera préférée aux autres variables. Le gain d'information est la différence entre l'incertitude à priori et l'incertitude à posteriori.
- Les critères de distance : Ces mesures s'intéressent au pouvoir discriminant d'une variable.
- Les critères d'indépendance : Ils regroupent toutes les mesures de corrélation ou d'association. Ils permettent de calculer le degré avec lequel une variable exogène est associée à une variable endogène.
- Les critères de consistance : Ils sont liés au biais des variables minimum (min-features bias). Ces mesures recherchent l'ensemble de variables le plus petit qui satisfait un pourcentage d'inconsistance minimum défini par l'utilisateur. Deux objets sont inconsistants si leurs modalités sont identiques et s'ils appartiennent à deux classes différentes. Ces mesures peuvent permettre de détecter les variables redondantes.
- Les critères de précision : Ils utilisent le classificateur comme fonction d'évaluation. Le classificateur choisit, parmi tous les sous-ensembles de variables, celui qui est à l'origine de la meilleure précision prédictive. Cette catégorie de critères est celle utilisée par toutes les méthodes enveloppe.

Du point de vue efficacité, la performance se caractérise par la capacité du modèle à prédire les défauts et les non-défauts. Pour cette étape, nous nous servons de la matrice de confusion pour notre spécification finale dans un premier temps. Par la suite, on se servira du processus d'estimation «in the sample» – prédiction «out of the sample» qui mettra à contribution également la matrice de confusion. La ROC curve et le test Hosmer-Lemeshow seront également mis à contribution. Nous reproduisons les grandes lignes de ces processus et mesures qui viendront soutenir la pertinence de notre spécification finale.

1.3.2.1. Le test de performance :

Qualité d'ajustement du modèle :

R^2 Ajusté : Il est compris en 0 et 1 ($0 < R^2 < 1$). Plus il est proche de 1, meilleur est le modèle.

1.3.2.2. La matrice de confusion :

La capacité d'un modèle à prédire les défauts et les non-défauts est donnée par la matrice de confusion. Les résultats de la classification des défauts et non défauts déterminés au moyen d'un seuil préétabli, sont comparés aux défauts et non défauts réalisés. Fondamentalement, il s'agit d'un tableau qui fournit le type de classification et le pourcentage associé à chaque type. Ces types sont définis en termes de bonnes classifications et de mauvaises classifications. Les classifications sont dites :

- Bonnes : lorsqu'un défaut prédit correspond à un défaut actuel (sensibilité du modèle) ou un non-défaut prédit correspond à un non-défaut actuel (spécificité du modèle).
- Mauvaises : lorsque les défauts prédits comme vrais correspondent à des non-défauts actuels. Cette mauvaise classification traduit une erreur de type II. Ce type d'erreur peut amener la banque à revoir son niveau de capital économique pour se prémunir d'éventuels chocs, ce qui se traduit par un manque à gagner et par ailleurs peut affecter la qualité de crédit du client. Inversement, une mauvaise classification qui classe les non-défauts prédits comme vrais alors qu'ils correspondent à des défauts actuels, traduit une erreur de type I. Les conséquences pour la banque d'une telle classification sont encore plus importantes puisqu'elles se matérialisent en termes de capital et d'intérêt.

De façon équivalente, nous pouvons définir les mauvaises classifications comme suit :
Si nous supposons une hypothèse nulle définie comme telle :

$H_0 : Y_i = 1$ (*être en défaut*). Les mauvaises classifications seront alors traduites ainsi :

- L'erreur de type I correspond au rejet de H_0 quand H_0 est vrai
- L'erreur de type II correspond à ne pas rejeter H_0 quand H_0 est fausse

Le tableau suivant correspond à une matrice de confusion générale :

		Réel	
		1	0
Prédiction	1	VP	FP
	0	FN	VN

1.3.2.3. La ROC curve :

La ROC curve (Receiver Operating Characteristic curve), du nom de courbe de spécificité en français, est une courbe qui permet de synthétiser l'information résultant de différentes matrices de confusions associées à des seuils de conservatismes correspondants. Elle correspond alors à une mesure de la performance d'un classificateur binaire. Son axe des abscisses correspond à (la mesure de spécificité du modèle), soit le pourcentage d'erreur de type II. La mesure de sensibilité est reportée sur son axe des ordonnées, soit le pourcentage de défauts correctement classés comme des défauts par le modèle. Elle est alors obtenue en faisant varier le seuil de conservatisme du modèle qui permet de générer différentes combinaisons bonnes de classification de défaut et d'erreurs de type II. Ces différentes combinaisons permettent ainsi de déterminer dans quelle mesure le modèle est précis pour discriminer les défauts et des non-défauts. L'aire sous la ROC curve du nom de mesure AUC (area under curve) est une mesure de la performance du modèle dans la prédiction du défaut actuel. Un modèle parfait aura une mesure AUC de 1. Ainsi, plus le modèle est précis, plus la courbure de la ROC curve est proche du coin gauche du graphique vers le haut et la mesure AUC est proche de 1.

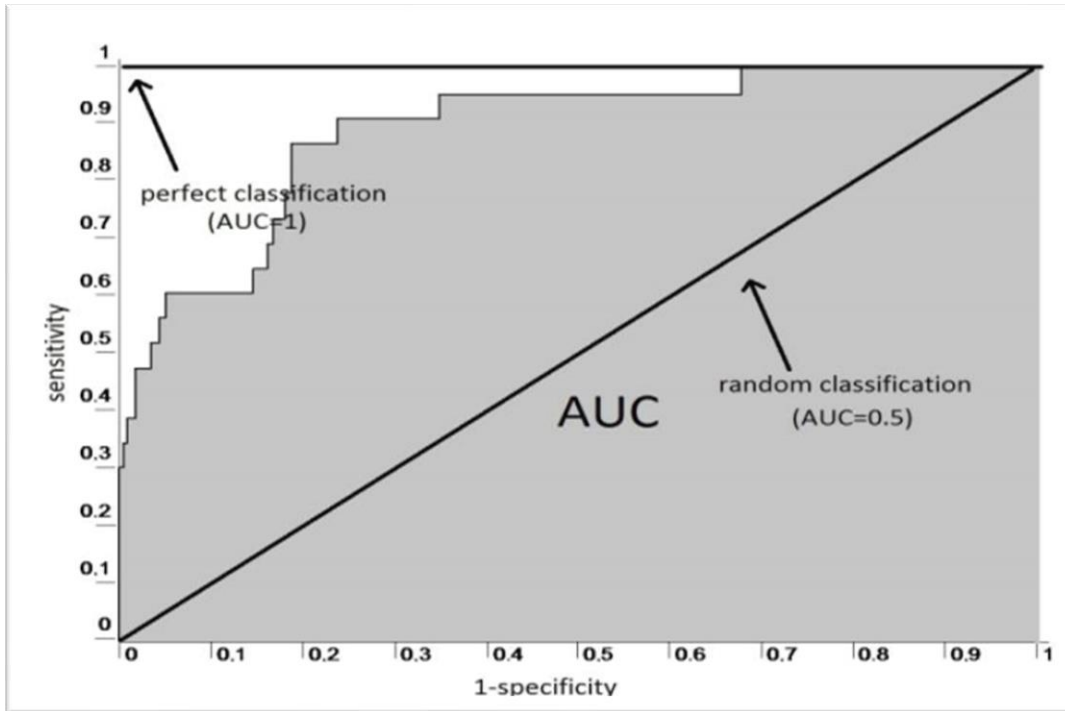


Figure 5 : ROC Curve

1.4. Adéquation du modèle :

Une fois le modèle construit, on peut déterminer sa qualité d'ajustement aux données.

✚ **Test de Hosmer et Lemershow** : il est basé sur un regroupement des probabilités prédites par le modèle. On calcule, pour chacun des groupes, le nombre observé de réponses positives $y=1$ et de réponses négatives $y = 0$, que l'on compare au nombre espéré prédit par le modèle. Une distance entre les fréquences observées et prédites au moyen d'une statistique de Khi-2 est alors calculée. Lorsque cette statistique est petite, on considère que le modèle est bien calibré.

Cependant, la puissance de ce test est relativement faible lorsque la taille de l'échantillon est inférieure à 400.

✚ **Test d'adéquation de la déviance** : Ce test permet de mesurer l'ajustement d'un modèle. Tout en sachant que le modèle saturé est le meilleur modèle en termes de qualité d'ajustement. Pour mesurer l'ajustement d'un modèle M_β , nous allons le comparer au modèle saturé en effectuant un test de rapport de vraisemblance (appelé déviance). Ce test n'est valable

qu'en présence de données répétées. Généralement, on écrit les hypothèses d'un test entre modèles emboîtés en termes de nullité de certains coefficients (ceux du grand modèle qui ne sont pas dans le petit). Il est difficile de présenter dans un cadre général une telle écriture puisque l'on a pas d'écriture générale de modèle saturé. C'est pourquoi, on commettra l'abus d'écrire les hypothèses sous cette forme :

H_0 : " le modèle considéré à M_β de dimension p est adéquat " (les données sont bien générées selon le modèle logistique en question) contre **H_1 : " le modèle saturé adéquat "**.

Cependant, dans la plupart des cas, on pourra écrire les hypothèses de ce test en termes de nullité de certains coefficients du modèle saturé.

Le test de la déviance compare le modèle saturé au modèle considéré au moyen de la déviance :

-Si la déviance est grande, alors le modèle considéré est loin du modèle saturé et que par conséquent il n'ajuste pas bien les données ;

-Par contre, si la déviance est proche de 0, le modèle considéré sera adéquat. Pour quantifier cette notion de "proche de 0" et de "grande déviance", la loi de la déviance sous H_0 (le modèle considéré est le vrai modèle) sera utile.

En effet si H_0 est vraie, le modèle considéré est vrai par définition. La déviance sera répartie sur R^+ mais avec plus de chance d'être proche de 0. Par contre si H_0 n'est pas vraie la déviance sera répartie sur $^+$ mais avec plus de chance d'être éloignée de 0. Il faut donc connaître la loi de la déviance sous H_0 . En présence de données répétées, si le nombre de répétitions n_t de chaque point tend vers ∞ , alors sous H_0 la déviance :

$$D_{M_\beta} = 2 (\mathcal{L}_{Sat} - \mathcal{L}_n(\hat{\beta}))$$

converge en loi vers un χ^2_{T-p} où p est la dimension de M_β .

Ainsi, on niveau α , on rejettera H_0 si la valeur observée de D_{M_β} est supérieure au quantile d'ordre $1 - \alpha$ de la loi χ^2_{T-p} .

Test d'adéquation de Pearson, Analyse des résidus et Effet de levier

2. Famille de modèles aléatoires : Bagging et Random Forest (les forêts aléatoires) :

Ces deux méthodes reposent sur une construction aléatoire d'une famille de modèles : bagging pour bootstrap aggregating (Breiman 1996) et les forêts aléatoires (random forests) de Breiman (2001) qui propose une amélioration du bagging spécifique aux modèles définis par des arbres binaires (CART). Ces algorithmes se sont développés à la frontière entre apprentissage machine (machine learning) et Statistique. Les principes du bagging ou du boosting s'appliquent à toute méthode de modélisation (régression, CART, réseaux de neurones) mais n'ont d'intérêt, et réduisent sensiblement l'erreur de prévision, que dans le cas de modèles instables, donc plutôt non linéaires. Ainsi, l'utilisation de ces algorithmes n'a guère de sens avec la régression multilinéaire ou l'analyse discriminante. Ils sont surtout mis en œuvre en association avec des arbres binaires comme modèles de base. En effet, l'instabilité déjà soulignée des arbres apparaît alors comme une propriété nécessaire à la réduction de la variance par agrégation de modèles.

2.1. Bagging : Principe et algorithme

Soit Y une variable à expliquer quantitative ou qualitative, X^1, \dots, X^p les variables explicatives et $f(x)$ un modèle fonction de $x = \{x^1, \dots, x^p\} \in \mathbb{R}^p$.

On note n le nombre d'observations et

$z = \{(x^1, y^1), \dots, (x^n, y^n)\}$ un échantillon de loi F .

Considérant B échantillons indépendants notés $\{Z_b\}_{b=1, B}$, une prévision par agrégation de modèles est définie ci-dessous dans le cas où la variable à expliquer Y est:

- quantitative : $\hat{f}_B(\cdot) = \frac{1}{B} \sum_{b=1}^B \hat{f}_{Z_b}(\cdot)$
- qualitative : $\hat{f}_B(\cdot) = \arg \max_j \text{card} \{b \mid \hat{f}_{Z_b}(\cdot) = j\}$

avec \hat{f}_{Z_b} : estimation du modèle sur l'échantillon b

Dans le premier cas, il s'agit d'une simple moyenne des résultats obtenus pour les modèles associés à chaque échantillon, dans le deuxième, un comité de modèles est constitué pour voter et élire la réponse la plus probable. Dans ce dernier cas, si le modèle retourne des probabilités associées à chaque modalité comme en régression logistique ou avec les arbres de décision, il est aussi simple de calculer des moyennes de ces probabilités.

Le principe est élémentaire, moyenner les prévisions de plusieurs modèles indépendants permet de réduire la variance et donc de réduire l'erreur de prévision.

Cependant, il n'est pas réaliste de considérer B échantillons indépendants. Cela nécessiterait généralement trop de données. Ces échantillons sont donc remplacés par B répliques d'échantillons bootstrap obtenus chacun par n tirages avec remise selon la mesure empirique \hat{F} . Ceci conduit à l'algorithme ci-dessous.

Algorithm 1 *Bagging*

Soit \mathbf{x}_0 à prévoir et

$\mathbf{z} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ un échantillon

for $b = 1$ à B **do**

Tirer un échantillon bootstrap \mathbf{z}_b^* .

Estimer $\hat{f}_{\mathbf{z}_b}(\mathbf{x}_0)$ sur l'échantillon bootstrap.

end for

Calculer l'estimation moyenne $\hat{f}_B(\mathbf{x}_0) = \frac{1}{B} \sum_{b=1}^B \hat{f}_{\mathbf{z}_b}(\mathbf{x}_0)$ ou le résultat du vote.

Figure 6 : Algorithme Bagging

Erreur out-of-bag :

Il est naturel et techniquement facile d'accompagner ce calcul par une estimation out-of-bag (o.o.b.) de l'erreur de prévision car sans biais, ou plutôt pessimiste, comme en validation croisée. C'est l'ensemble des exemples qui ne sont pas sélectionnés dans les échantillons

bootstrap. C'est un paramètre qui permet l'évaluation interne du classifieur et l'estimation de l'importance des variables pour la sélection de variables

Pour chaque observation (y_i, x_i) considérer les seuls modèles estimés sur un échantillon bootstrap ne contenant pas cette observation (à peu près 1/3). Prévoir la valeur y_b comme précédemment (moyenne ou vote), calculer l'erreur de prévision associée et moyenner sur toutes les observations.

En pratique, CART (*Classification And Regression Trees*) est souvent utilisée comme méthode de base pour construire une famille de modèles c'est-à-dire d'arbres binaires. L'effet obtenu, par moyennage d'arbres, est une forme de "lissage" du pavage de l'espace des observations pour la construction des règles de décision. Trois stratégies d'élagage sont possibles :

1. laisser construire et garder un arbre complet pour chacun des échantillons en limitant le nombre minimale (5 par défaut) d'observation par feuille ;
2. construire un arbre d'au plus q feuilles ou de profondeur au plus q ;
3. construire à chaque fois l'arbre complet puis l'élaguer par validation croisée.

La première stratégie semble en pratique un bon compromis entre volume des calculs et qualité de prévision. Chaque arbre est alors affecté d'un faible biais et d'une grande variance mais la moyenne des arbres réduit avantageusement celle-ci. En revanche, l'élagage par validation croisée pénalise les calculs sans, en pratique, gain substantiel de qualité. Cet algorithme a l'avantage de la simplicité, il s'adapte et se programme facilement quelle que soit la méthode de modélisation mise en œuvre.

Il pose néanmoins quelques problèmes :

- temps de calcul pour évaluer un nombre suffisant d'arbres jusqu'à ce que l'erreur de prévision out-of-bag ou sur un échantillon validation se stabilise et arrête si elle tend à augmenter ;

- nécessiter de stocker tous les modèles de la combinaison afin de pouvoir utiliser cet outil de prévision sur d'autres données,
- l'amélioration de la qualité de prévision se fait au détriment de l'interprétabilité.

Le modèle finalement obtenu devient une boîte noire.

2.2. Forêts aléatoires :

Dans le cas spécifique des modèles d'arbres binaires de décision (CART), Breiman (2001) propose une amélioration du bagging par l'ajout d'une composante aléatoire. L'objectif est donc de rendre plus indépendants les arbres de l'agrégation en ajoutant du hasard dans le choix des variables qui interviennent dans les modèles.

Les Forêts Aléatoires ou Random Forest sont parmi les méthodes de Machine Learning les plus populaires grâce à leur précision, leur robustesse et leur facilité d'utilisation.

Néanmoins elle peut conduire aussi à de mauvais résultats notamment lorsque le problème sous-jacent est linéaire et donc qu'une simple régression PLS conduit à de bonnes prévisions même en grande dimension.

Plus précisément, la variance de la moyenne de B variables indépendantes, identiquement distribuées, chacune de variance σ^2 , est σ^2/B .

Si ces variables sont identiquement distribuées et en supposant qu'elles sont corrélées deux à deux de corrélation ρ , Breiman (2001) montre que la variance de la moyenne devient :

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

Comme dans le cas indépendant, le 2ème terme décroît avec B mais le premier terme limite considérablement l'avantage du bagging si la corrélation est élevée. C'est ce qui motive principalement la randomisation introduite dans l'algorithme ci-dessous afin de réduire ρ entre les prévisions fournies par chaque modèle.

Le bagging est appliqué à des arbres binaires de décision en ajoutant un tirage aléatoire de m variables explicatives parmi les p .

Algorithm 2 Forêts Aléatoires

Soit \mathbf{x}_0 à prévoir et

$\mathbf{z} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ un échantillon

for $b = 1$ à B **do**

Tirer un échantillon bootstrap \mathbf{z}_b^*

Estimer un arbre sur cet échantillon avec **randomisation** des variables : la recherche de chaque division optimale est précédée d'un tirage aléatoire d'un sous-ensemble de m prédicteurs.

end for

Calculer l'estimation moyenne $\hat{f}_B(\mathbf{x}_0) = \frac{1}{B} \sum_{b=1}^B \hat{f}_{\mathbf{z}_b}(\mathbf{x}_0)$ ou le résultat du vote.

Figure 7 : Algorithme Random Forest

Le principal avantage de cet algorithme est qu'il permet d'éviter le danger que représente le sur-apprentissage pour toute méthode de prédiction basée sur l'induction. BREIMAN (2001) démontre que lorsque le nombre d'arbres impliqués dans la forêt de prédiction augmente, le taux d'erreur en généralisation converge vers une valeur limite, dont une borne supérieure peut être estimée sur une base des caractéristiques intrinsèques de la forêt. Il s'agit de la propriété de convergence des forêts aléatoires. Cela explique pourquoi les forêts aléatoires ne font pas de sur-apprentissage lorsque le nombre d'arbres de la forêt augmente mais plutôt convergent vers une valeur limite de l'erreur OOB.

Erreur OOB : tous les échantillons OOB sont évalués par l'arbre et l'erreur mesurée. Ensuite on permute aléatoirement les valeurs sur chaque attribut j et on mesure le taux d'erreur à nouveau. La valeur finale est la dégradation moyenne (changement du taux d'erreurs) sur tous les arbres.

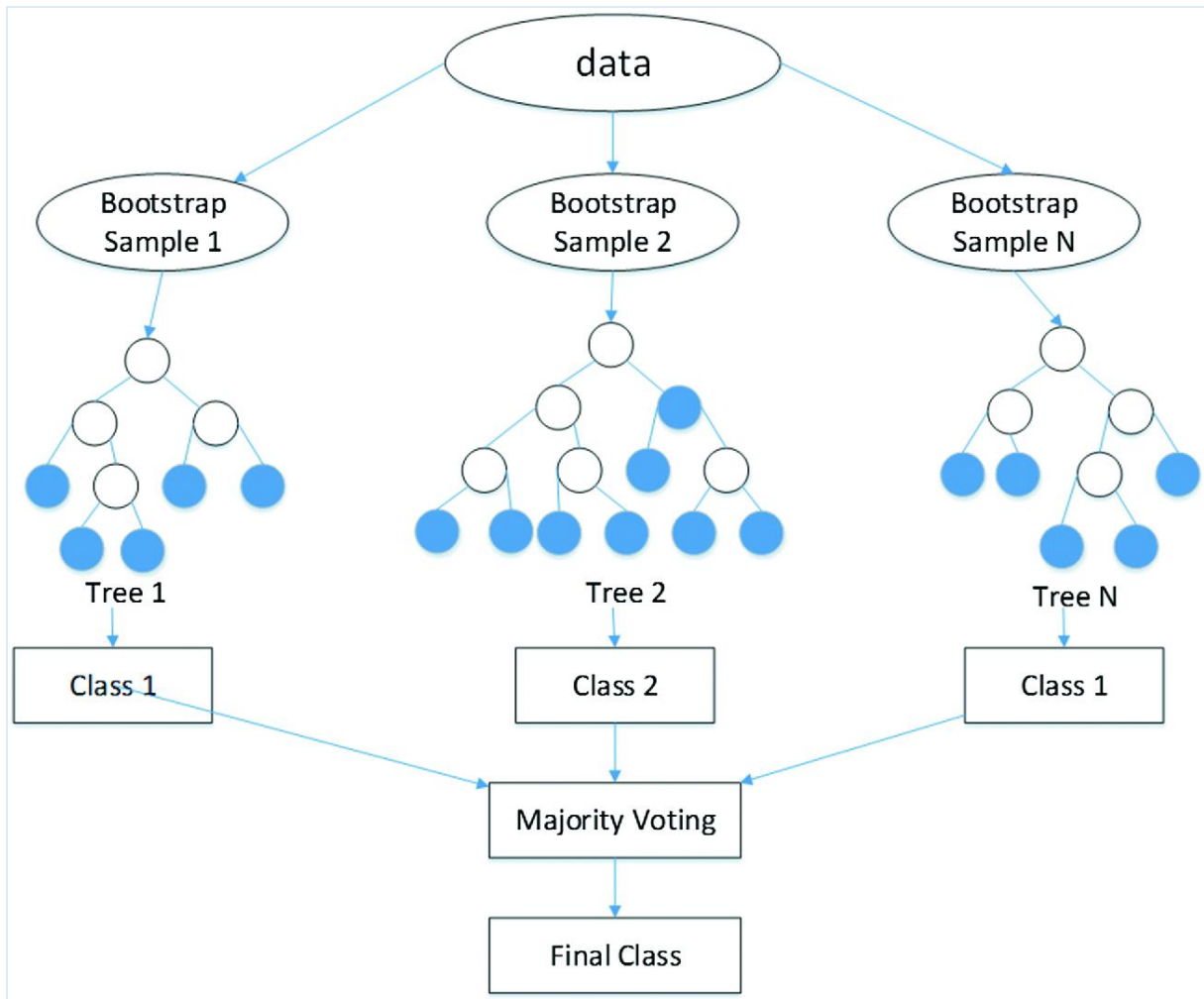


Figure 8 : Illustration Random Forest

L'idée est de créer un grand nombre d'arbres de décisions de façon aléatoires, à partir de différents sous-ensembles de données de l'ensemble de données initial. Le fait de considérer différents sous-ensembles est important : cela réduit les risques d'erreur, puisque nos arbres seront peu corrélés. On évite alors également le problème du surapprentissage, qui intervient lorsque l'arbre construit s'est trop adapté à l'échantillon considéré.

Il aura considéré tous les tests possibles, chaque feuille ne représentant qu'un unique candidat. Cet arbre sera donc fiable à 100% pour l'échantillon ayant permis sa construction (puisque'il prend en compte tous les tests possibles), mais ne sera pas généralisable à d'autres échantillons.

Le candidat est alors testé sur chacun de ces arbres (qui peuvent être plus d'une centaine). L'intérêt de la forêt est de procéder par vote majoritaire quant aux résultats obtenus.

On réduit ainsi la marge d'erreur que peut avoir un arbre seul. Plus l'on dispose d'arbres, plus la forêt sera fiable.

2.3. Les critères d'évaluation :

Les performances du système de classification sont évaluées en utilisant la matrice de confusion qui est un outil performant et bien adapté aux problèmes de classification.

Matrice de confusion :

		Réel	
		1	0
Prédiction	1	VP	FP
	0	FN	VN

Figure 9 : Matrice de confusion

Les lignes de la matrice de confusion représentent les prédictions alors que les colonnes représentent les classes réelles. Le calcul des vrais positifs (VP), des vrais négatifs (VN), des faux positifs (FP) et des faux négatifs (FN), le pourcentage de sensibilité (S_e), la spécificité (S_p) et le taux de classification (T_c) permet de faire cette évaluation. Leurs définitions respectives sont les suivantes :

- VP : Vrai Positif : nombre de positifs classés positifs
- VN : Vrai Négatif : nombre de négatifs classés négatifs
- FP : Faux Positif : nombre de négatifs classés positifs
- FN : Faux négatif : nombre de positifs classés négatifs

Ils permettent de calculer les termes suivants :

La sensibilité : C'est le taux de vrais positifs qui se calcule comme suit.

$$S_e = \frac{VP}{VP + VN} = TVP$$

La spécificité :

$$S_p = \frac{VN}{VN + FP} = 1 - TFP$$

Le taux de classification : c'est le pourcentage des exemples correctement classés, il est calculé par :

$$T_c = \frac{VP + VN}{VP + VN + FP + FN}$$

Cependant, la matrice de confusion peut être réduite à 2 taux indépendants de la distribution des classes : TVP ou Sensibilité et le TFP.

3. Comparaison des deux techniques de rating :

Le tableau suivant résume les avantages et les inconvénients de chacune des méthodes à appliquer.

Techniques	Avantages	Inconvénients
Régression logistique	<ul style="list-style-type: none"> -Ne nécessite pas de relation linéaire entre les variables dépendantes et indépendantes -Peut gérer différents types de relations -Le modèle de régression logistique agit non seulement comme un modèle de classification, mais donne également des probabilités -La régression logistique donne en plus d'une mesure de la pertinence d'un prédicteur (taille du coefficient), son sens d'association (positif ou négatif) -Elle s'avère très efficace lorsque l'ensemble de données a des caractéristiques linéairement séparables. 	<ul style="list-style-type: none"> -Nécessite des échantillons de grande taille (car les estimations du maximum de vraisemblance sont moins puissantes pour les échantillons de faible taille) -Les variables indépendantes ne doivent pas être corrélées les unes avec les autres -La régression logistique est moins sujette au surajustement, mais elle peut être sur-ajustée dans des ensembles de données de grande dimension. -Il est difficile de saisir des relations complexes en utilisant la régression logistique (des algorithmes plus puissants et complexes tels que les réseaux de neurones peuvent facilement surpasser cet algorithme) -Ne traite pas les valeurs manquantes.
Random Forest	<ul style="list-style-type: none"> -Random forest présente l'avantage d'utiliser plus intelligemment l'ensemble de ses données initiales, afin de limiter ses erreurs. Toutefois, l'algorithme ne pourra jamais être à 100% fiable (il faudrait disposer d'une infinité d'arbres) -Bonnes performances en prédiction -Paramétrage simple (B et m) -Pas de problème d'overfitting (augmenter B) -Mesure de l'importance des variables -Evaluation de l'erreur intégrée (OOB) -Possibilité de programmation parallèle -Automatisation des valeurs manquantes présentes dans les données -La normalisation des données n'est pas nécessaire car elle utilise une approche basée sur des règles. 	<ul style="list-style-type: none"> -Problème si le nombre de variables pertinentes est très faibles, dans l'absolu et relativement au nombre total de variables -Déploiement d'un tel modèle reste compliqué -Il nécessite beaucoup de puissance de calcul ainsi que des ressources car il construit de nombreux arbres pour combiner leurs sorties -Il nécessite également énormément de temps pour la formation (car il combine de nombreux arbres de décision pour déterminer la classe) -En raison de l'ensemble des arbres de décision, il souffre également d'interprétabilité et ne parvient pas à déterminer la signification de chaque variable.

Tableau 3 : Comparaison entre la régression logistique et la méthode Random Forest

Chapitre 3 : Elaboration d'un modèle de notation interne

– Partie Pratique

I. Données et statistique descriptive des variables :

1. Analyse de la base de données :

L'entreprise pour laquelle nous devons construire un modèle de notation interne est une association à but non-lucratif du groupe Banque Populaire qui a pour mission de distribuer des micro-crédits, afin de permettre à des personnes économiquement faibles de créer ou de développer leur propre activité de production ou de service et d'assurer leur insertion économique. Aussi, elle a pour but d'effectuer au profit de ses clients, toutes opérations connexes liées à l'octroi de micro-crédits, notamment la formation, le conseil et l'assistance technique.

La base de données sur laquelle portera notre étude contient des informations sur **627** clients relatives à **16 variables** (10 quantitatives et 6 qualitatives) dont les caractéristiques sont les suivantes :

Variable	Type
Statut matrimonial	Qualitative
Nombre de personnes à charge	Quantitative
Age de l'emprunteur	Quantitative
Historique de l'emprunteur au sein de la Fondation (Nombre de prêts remboursés)	Quantitative
Propriété de business à financer	Qualitative
Forme juridique	Qualitative
Durée de l'activité en mois	Quantitative
Valeur du collatéral/Montant de l'emprunt (Taux de couverture)	Quantitative
Valeur des actifs de l'entreprise / Montant de l'emprunt	Quantitative
Nombre d'incidents des crédits en cours	Quantitative

Capacité de remboursement (Mensualités du prêt / Revenu net corrigé)	Quantitative
Garantie des sources de revenu	Qualitative
Formalité de l'entreprise	Qualitative
Récentes requêtes sur le dossier de crédit	Quantitative
Lieu d'exercice de l'activité	Qualitative
Nombre d'années dans le business (pour activités similaires)	Quantitative

Tableau 4 : Les variables et leurs types

Répartition des variables qualitatives :

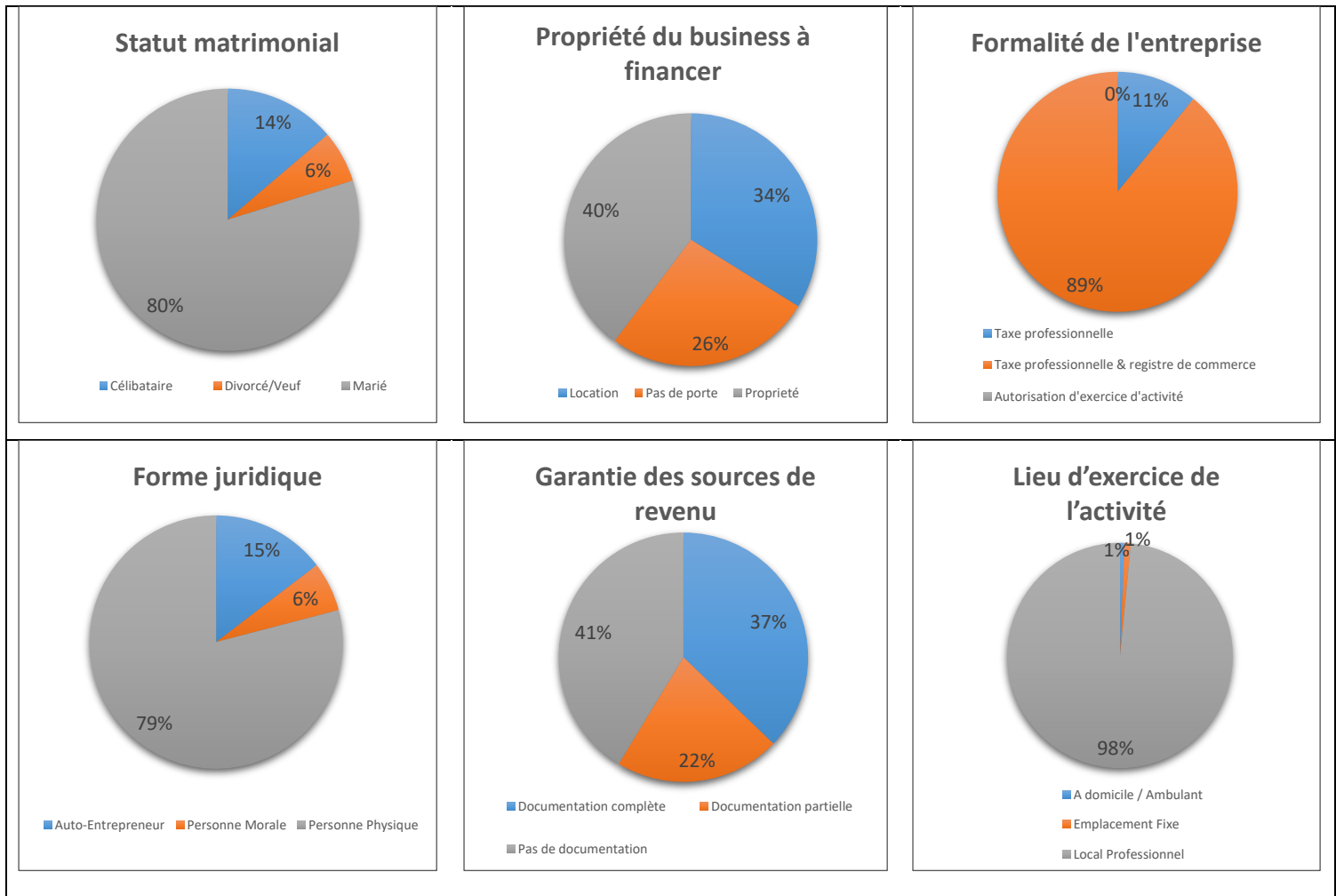


Figure 10 : La répartition des variables qualitatives

Hormis les variables « Lieu d'exercice de l'activité » et « Formalité de l'entreprise », les clients sont plus ou moins bien répartis entre les différentes modalités des variables. C'est-

à-dire qu'il n'y a pas de concentration sur une seule modalité ou qu'aucune modalité n'est inexistante.

Distribution des variables quantitatives :

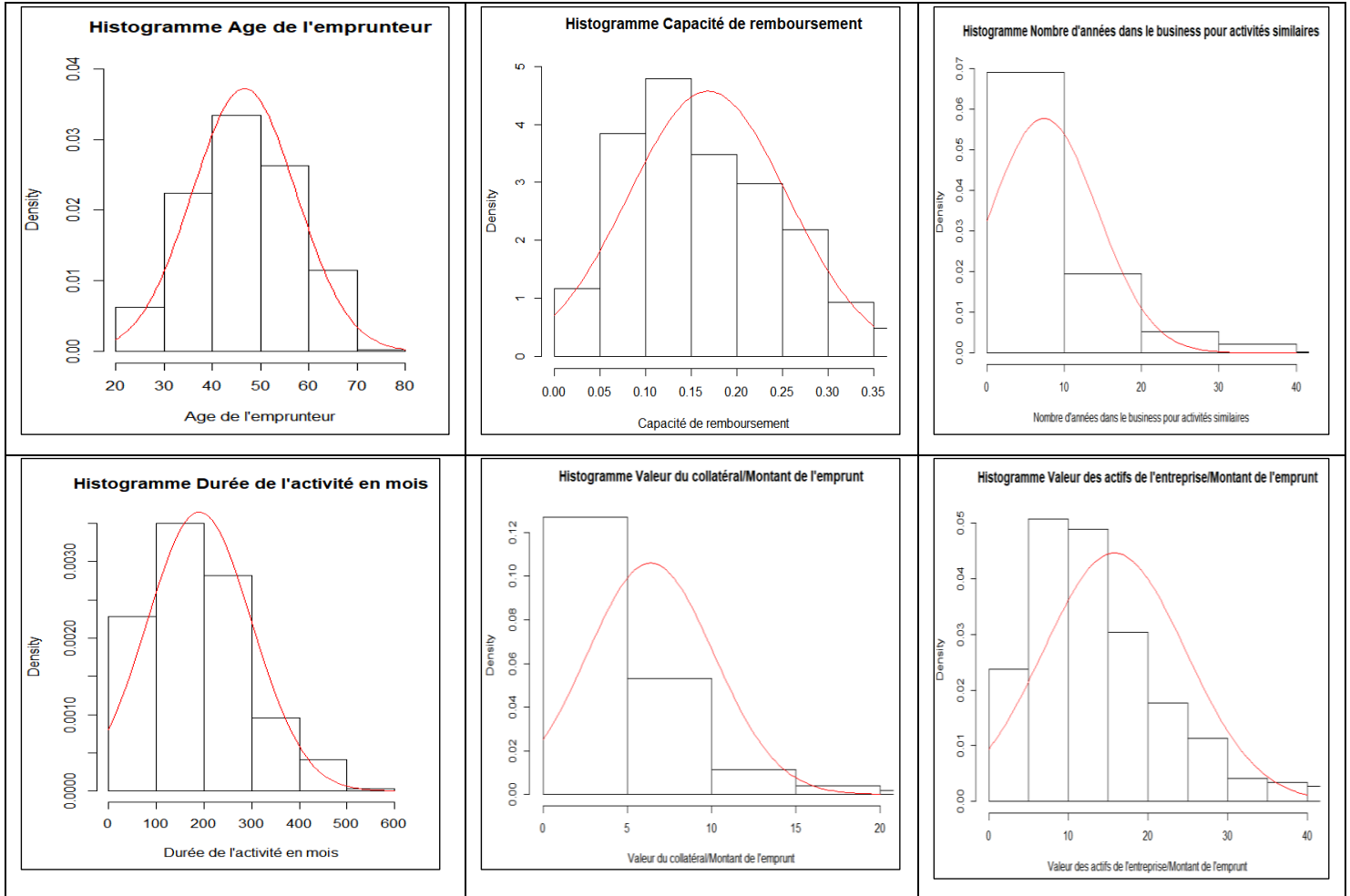


Figure 11 : La distribution des variables quantitatives

2. Traitement de la base et analyse descriptive des données :

2.1. Valeurs manquantes :

Les données manquantes sont fréquemment rencontrées dans les évaluations économiques. Les ignorer peut entraîner outre une perte de précision, de forts biais dans les

analyses. Comme le montre le tableau ci - après, notre base de données contient un très grand nombre de valeurs manquantes.

Variable	Nombre de valeurs manquantes
Statut matrimonial	0
Nombre de personnes à charge	3
Age de l'emprunteur	0
Historique de l'emprunteur au sein de la Fondation (Nombre de prêts remboursés)	9
Propriété du business à financer	1
Forme juridique	2
Durée de l'activité en mois	0
Valeur du collatéral/Montant de l'emprunt (Taux de couverture)	0
Valeur des actifs de l'entreprise / Montant de l'emprunt	0
Nombre d'incidents des crédits en cours	46
Capacité de remboursement (Mensualités du prêt / Revenu net corrigé)	0
Garantie des sources de revenu	54
Formalité de l'entreprise	120
Récentes requêtes sur le dossier de crédit	64
Lieu d'exercice de l'activité	17
Nombre d'années dans le business (pour activités similaires)	0

Tableau 5 : Valeurs manquantes

Traitement des valeurs manquantes :

Afin de conserver le maximum de données et d'éviter des écarts considérables, nous nous sommes penchés sur la simulation des valeurs manquantes. Cette simulation se base exclusivement sur la distribution des données de chacune des variables concernées.

Pour une meilleure illustration du procédé adopté, prenons en exemple la variable « Garantie des sources de revenu » qui totalise *54 valeurs manquantes* et analysons les différentes distributions avant et après simulation.

Cette variable a 3 modalités :

- Documentation complète
- Documentation partielle
- Pas de documentation

Le tableau suivant résume la répartition de ces modalités, d'une part en ne considérant pas les valeurs manquantes et d'autre part la répartition des données obtenues par simulation (54).

Répartition avant simulation		➔	Répartition valeurs simulées	
Documentation complète	40%		Documentation complète	43%
Documentation partielle	24%		Documentation partielle	24%
Pas de documentation	36%		Pas de documentation	33%
Total	573		Total	54

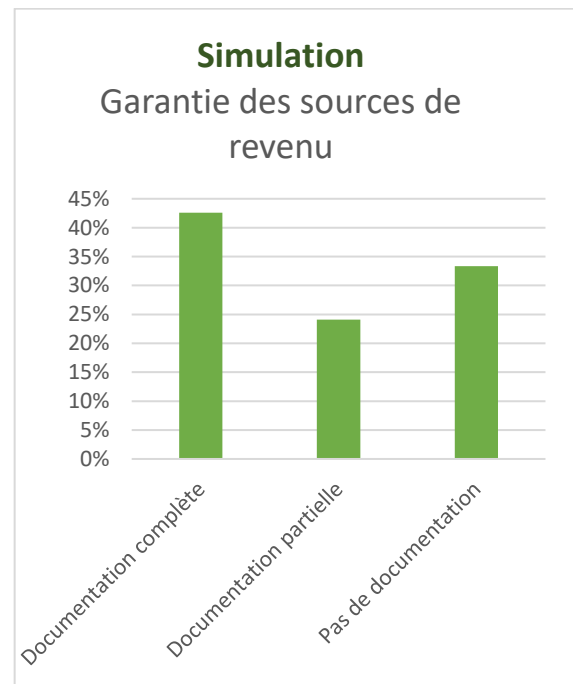
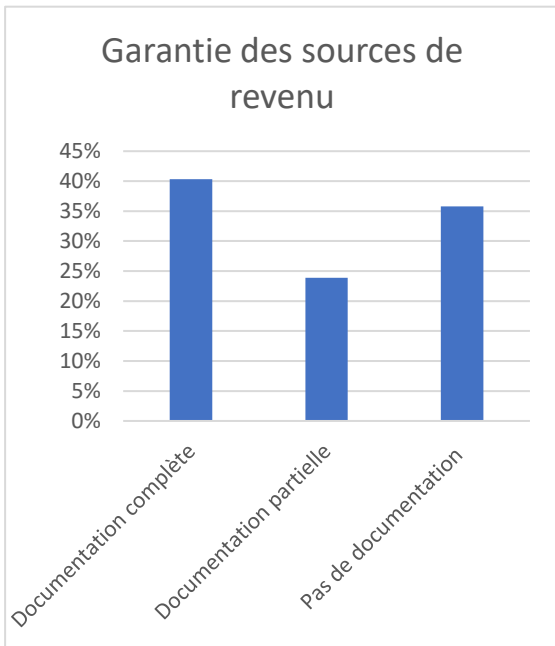


Figure 12 : Technique de traitement des missing values

Nous pouvons remarquer que les distributions sont assez semblables. Ce qui nous conforte dans le choix de la méthode adoptée pour palier au problème de missing values.

Cette méthode a été appliquée pour chaque variable où on observe des données manquantes. Ce qui nous a permis d'obtenir, sans perte d'informations une base de données sans valeurs manquantes.

2.2. Valeurs aberrantes :

Une valeur aberrante est une valeur extrême, anormalement différente de la distribution d'une variable. En d'autres termes, la valeur de cette observation diffère grandement des autres valeurs de la même variable : anormalement faible ou élevée.

Plusieurs algorithmes de Machine Learning sont sensibles aux données d'entraînement ainsi qu'à leurs distributions. Avoir des Outliers dans Training Set d'un algorithme de Machine Learning peut rendre la phase d'entraînement plus longue. Sans mentionner que l'apprentissage sera biaisé. Par conséquent, le modèle prédictif produit ne sera pas performant, ou du moins, loin d'être optimal.

Bien avant la phase d'apprentissage, les valeurs aberrantes influencent certains paramètres statistiques, comme la moyenne. Cela peut fausser notre compréhension du jeu de données et nous conduire à émettre des hypothèses erronées sur ce dernier. Détecter ses Outliers nous permettra de faire des suppositions plus aguerries.

On peut repérer les valeurs aberrantes en utilisant les boîtes à moustache. Il s'agit de la méthode la plus simple. Les Box Plot (Boîtes à Moustaches) permettent de visualiser la distribution d'une seule variable. Ces graphiques se basent sur la médiane, ainsi que les quartiles inférieurs et supérieurs Q_1 et Q_3 respectivement. Une valeur est considérée comme aberrante si la valeur absolue de l'écart avec Q_1 ou Q_3 est supérieure à plus de $1,5 \times$ Ecart interquartile. Plus précisément, une valeur aberrante est dite :

- ✓ Faible si elle est inférieure à : $Q_1 - 1,5 * IQ$
- ✓ Elevée si elle est supérieure à : $Q_3 + 1,5 * IQ$

Avec : $IQ = Q_3 - Q_1$: l'écart interquartile

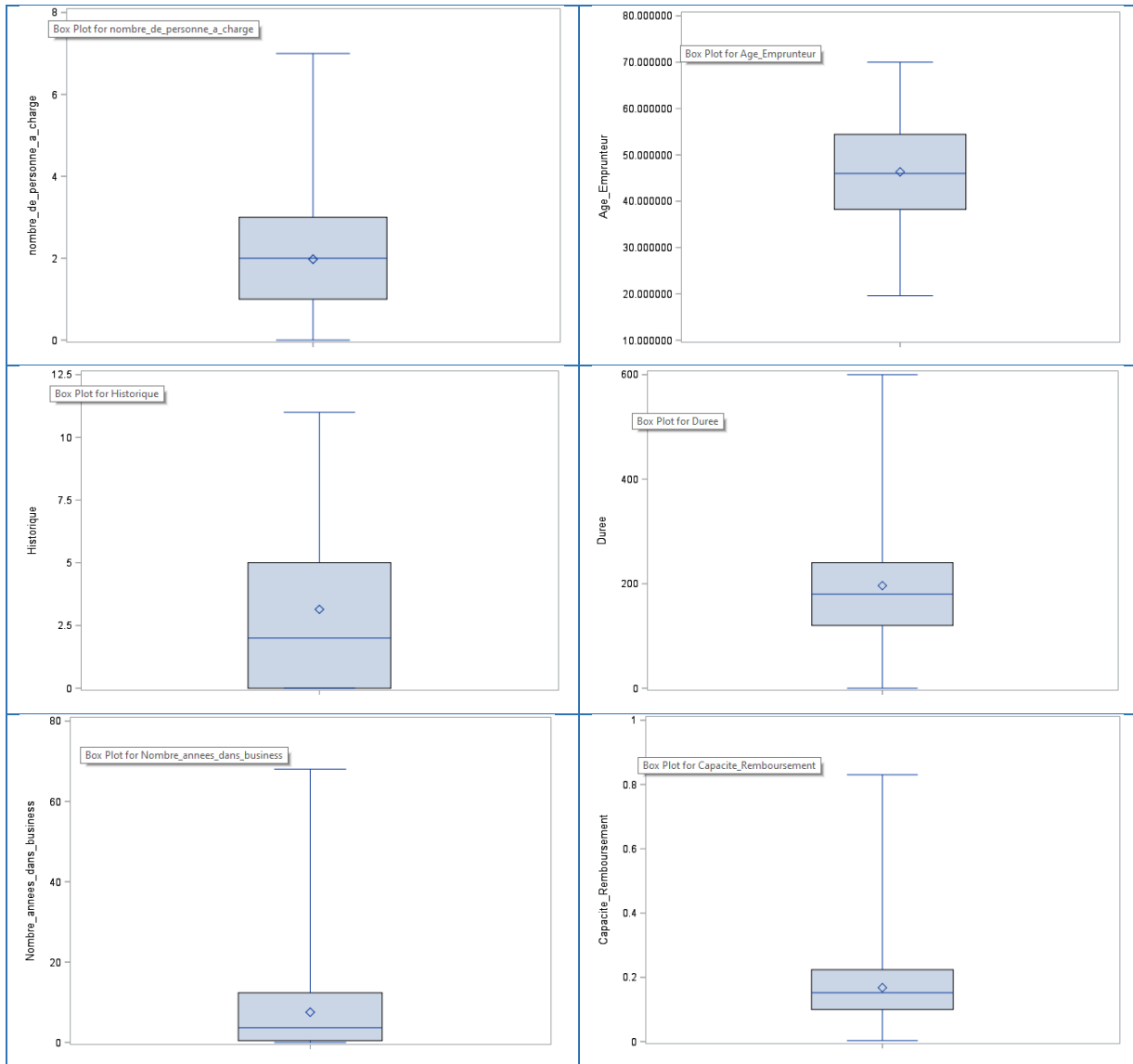


Figure 13 : Boxplots des variables qualitatives

Comme nous le constatons à travers ces différentes boîtes à moustaches tracées, notre base de données ne contient heureusement pas de valeurs extrêmes.

2.3. Répartition des clients sains et en défaut :

Est considéré comme client présentant un défaut, tout client ayant accusé un retard de paiement de plus de 30 jours (impayé > 30 jours).

L'identification des clients en défaut s'est faite par leur recherche dans la base des impayés (> 30 jours).

A chaque client trouvé dans cette base, est affecté une valeur 1 à la variable **défaut**.

Defaut				
Defaut	Fréquence	Pourcentage	Fréquence cumulée	Pctage cumulé
0	600	95.69	600	95.69
1	27	4.31	627	100.00

Figure 14 : Effectifs des clients sains et en défaut

Parmi les 627 clients recensés dans la base, 95,69 % d'entre eux sont de bons emprunteurs et par conséquent, le reste (4,31%) des clients ont un défaut de remboursement de crédit.

2.4. Analyse des corrélations :

Variables quantitatives :

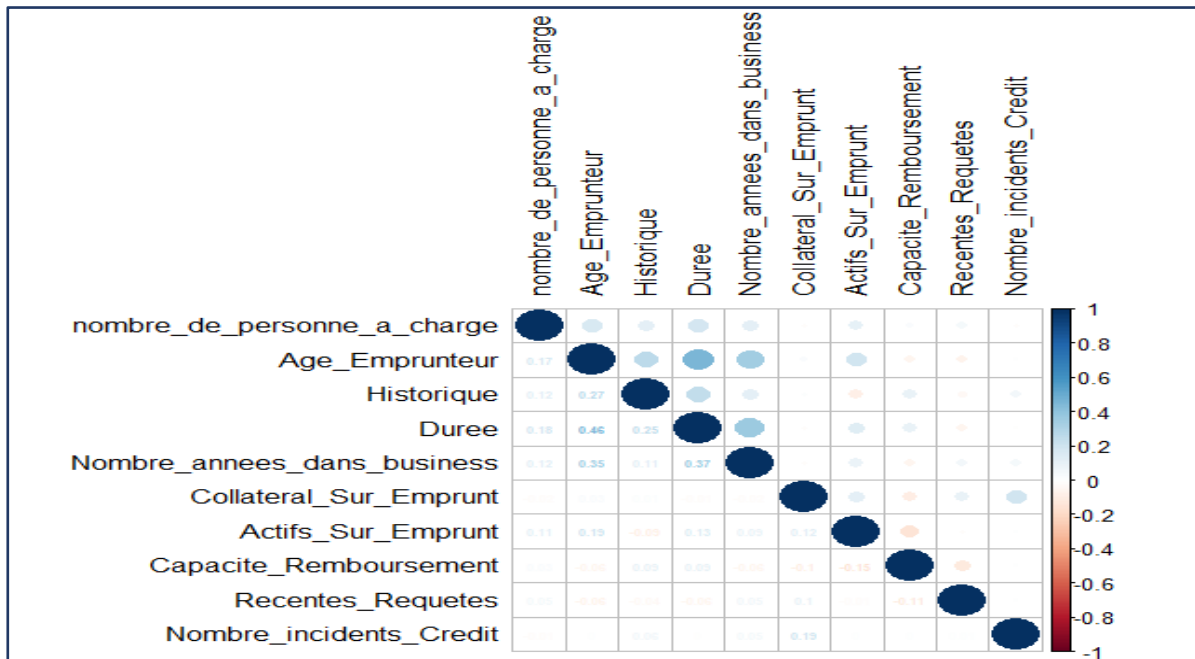


Figure 15 : Corrélation entre les variables quantitatives

Les variables quantitatives les plus corrélées sont :

- ✚ Durée de l'activité et Age de l'emprunteur
- ✚ Nombre d'années dans le business et Age de l'emprunteur
- ✚ Nombre d'années dans le business et Durée de l'activité

Variables quantitatives Vs Variables qualitatives :

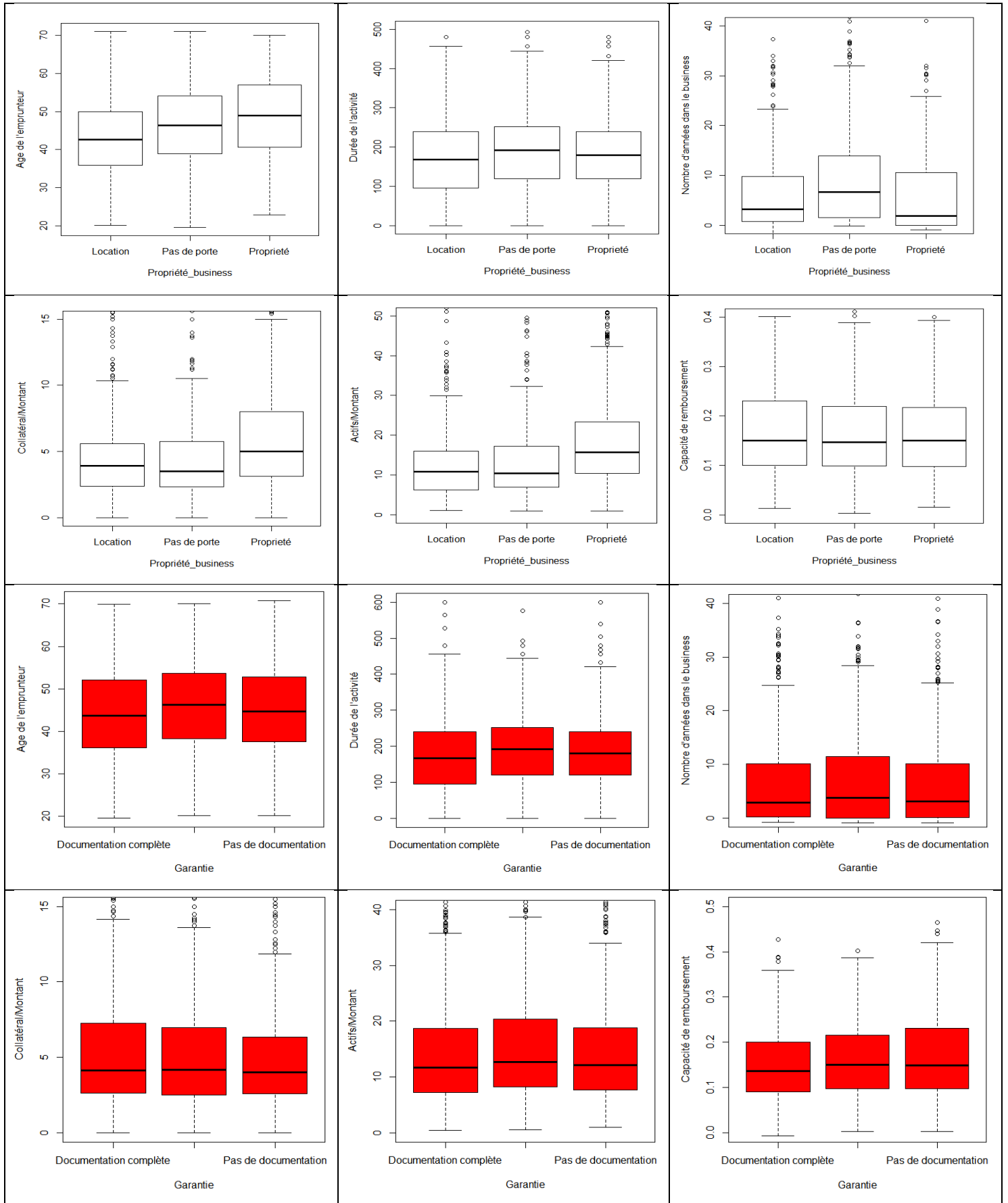


Figure 16 : Corrélation variables quantitatives Vs variables qualitatives

Récapitulatif des variables corrélées :

Le tableau suivant présente un résumé des corrélations entre les différentes variables choisis pour notre étude. Les cases comportant des disques « rouges » sont relatives aux variables fortement corrélées et celles contenant des disques « marron » font référence aux variables présentant une corrélation moyenne. Les cases vides quant à elles désignent des variables qui ne sont pas corrélées.

	Statut matrimonial	Nombre de personnes à charge	Age de l'emprunteur	Historique de l'emprunteur au sein de la fondation	Propriété du business à financer	Durée de l'activité en mois	Nombre d'années dans le business	Valeur du collatéral/montant de l'emprunt	Valeur des actifs de l'entreprise/montant de l'emprunt	Capacité de remboursement	Garantie des sources de revenu
Statut matrimonial											
Nombre de personnes à charge	●										
Age de l'emprunteur											
Historique de l'emprunteur au sein de la fondation	●		●								
Propriété du business à financer	●		●	●				●			
Durée de l'activité en mois			●	●							
Nombre d'années dans le business			●		●	●					
Valeur des actifs de l'entreprise/montant de l'emprunt					●						
Capacité de remboursement											
Garantie des sources de revenu	●	●	●	●	●		●				
Récentes requêtes sur le dossier de crédit					●						●

●	Fortement corrélées	●	Moyennement corrélées
---	---------------------	---	-----------------------

Tableau 6 : Récapitulatif des corrélations

II. Elaboration du modèle :

1. Régression logistique :

Nous appliquerons une régression logistique car notre variable dépendante (Défaut) est binaire : prend la valeur 1 si le client a connu un défaut et 0 sinon. Nous avons utilisé le logiciel SAS pour notre étude.

Tout d'abord, le test Global de d'hypothèse nulle $BETA = 0$ ($\beta = 0$) présente une p-value ($Pr > KHI2$) inférieur à 0.05. Ce qui signifie qu'au moins un des facteurs étudiés impact la probabilité de défaut.

Test de l'hypothèse nulle globale : BETA=0			
Test	Khi-2	DDL	Pr > Khi-2
Rapport de vrais	111.9758	21	<.0001
Score	126.0114	21	<.0001
Wald	45.7493	21	0.0014

Figure 17 : Test de nullité globale

1.1. Sélection des variables :

1.1.1. Par la méthode backward :

L'estimation des paramètres du modèle par la méthode du maximum de vraisemblance donne comme résultat :

Estimations par l'analyse du maximum de vraisemblance						
Paramètre		DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > Khi-2
Intercept		1	-5.1101	0.9510	28.8760	<.0001
Duree		1	-0.00738	0.00304	5.9153	0.0150
Capacite_Rembourseme		1	14.3563	2.9369	23.8952	<.0001
Nombre_incidents_Cre		1	1.0484	0.3209	10.6763	0.0011
Statut_Matrimonial	Divorcé/Veuf	1	-3.0694	1.8151	2.8595	0.0908
Statut_Matrimonial	Marié	1	-1.9381	0.5502	12.4074	0.0004
Propriete	Pas de porte	1	0.6871	0.5410	1.6132	0.2040
Propriete	Propriété	1	-2.4871	0.9672	6.6127	0.0101
Garantie	Documentation partielle	1	1.6755	0.8659	3.7446	0.0530
Garantie	Pas de documentation	1	2.1523	0.7396	8.4680	0.0036

Figure 18 : Estimation des paramètres du modèle issu de la régression logistique

Estimations des rapports de cotes			
Effet	Valeur estimée du point	95% Intervalle de confiance de Wald	
Duree	0.993	0.987	0.999
Capacite_Rembourseme	>999.999	>999.999	>999.999
Nombre_incidents_Cre	2.853	1.521	5.351
Statut_Matrimonial Divorcé/Veuf vs Célibataire	0.046	0.001	1.629
Statut_Matrimonial Marié vs Célibataire	0.144	0.049	0.423
Propriete Pas de porte vs Location	1.988	0.689	5.740
Propriete Propriété vs Location	0.083	0.012	0.554
Garantie Documentation partielle vs Documentation complète	5.342	0.979	29.154
Garantie Pas de documentation vs Documentation complète	8.605	2.019	36.670

Figure 19 : Odds ratio

Récapitulatif sur l'élimination en arrière						
Etape	Effet supprimé	DDL	Nombre dans	Khi-2 de Wald	Pr > Khi-2	Libellé de variable
1	Lieu_Exercice	2	15	0.0007	0.9997	Lieu_Exercice
2	Recentes_Requetes	1	14	0.0081	0.9281	Recentes_Requetes
3	Forme_Juridique	2	13	0.1784	0.9147	Forme_Juridique
4	Collateral_Sur_Empru	1	12	0.0221	0.8819	Collateral_Sur_Emprunt
5	nombre_de_personne_a	1	11	0.1446	0.7038	nombre_de_personne_a_charge
6	Historique	1	10	0.1621	0.6872	Historique
7	Formalite	1	9	0.2122	0.6450	Formalite
8	Nombre_annees_dans_b	1	8	0.9942	0.3187	Nombre_annees_dans_business
9	Age_Emprunteur	1	7	3.0891	0.0788	Age_Emprunteur
10	Actifs_Sur_Emprunt	1	6	3.0280	0.0818	Actifs_Sur_Emprunt

Figure 20 : Récapitulatif de la méthode de sélection Backward

La courbe ROC est un outil d'évaluation et de comparaison des modèles. C'est un outil graphique qui permet de visualiser les performances. Normalement, un seul coup d'œil doit permettre de voir le(s) modèle(s) susceptible(s) de nous intéresser.

La figure suivante présente les courbes ROC pour les différentes étapes lors de la création du modèle (méthode backward). Le modèle dominant est le modèle dont la courbe est « au-dessus » de toutes les autres courbes dans l'espace ROC.

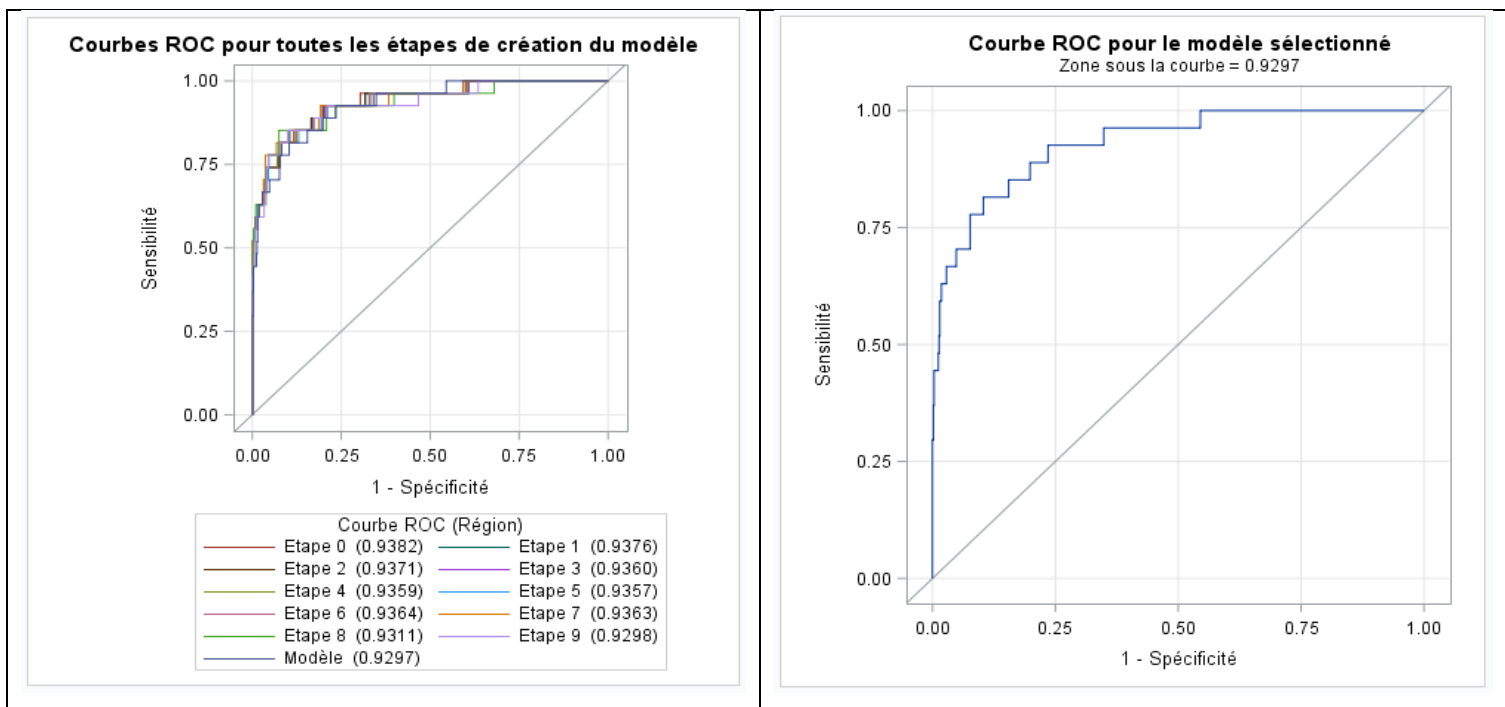


Figure 21 : Courbe ROC Backward

Même si la courbe ROC nous donne une représentation visuelle de la capacité discriminante, un indice numérique peut s'avérer utile pour juger la performance. Plusieurs indices ont été développés, mais un choix logique et simple est l'aire sous la courbe ROC : AUC.

AUC comprise entre...	Interprétation
0.5-0.6	Aucune relation
0.6-0.7	Liaison faible
0.7-0.8	Liaison significative
0.8-0.9	Liaison forte
0.9-1	Corrélation

Tableau 7 : Interprétation AUC en fonction des valeurs prises

Partition pour les tests de Hosmer et de Lemeshow					
Groupe	Total	Defaut = 1		Defaut = 0	
		Observé	Attendu	Observé	Attendu
1	63	0	0.01	63	62.99
2	63	0	0.02	63	62.98
3	64	0	0.05	64	63.95
4	63	0	0.10	63	62.90
5	63	1	0.20	62	62.80
6	63	0	0.38	63	62.62
7	64	1	0.73	63	63.27
8	63	2	1.56	61	61.44
9	63	4	3.95	59	59.05
10	58	19	20.00	39	38.00

Test d'adéquation de Hosmer et de Lemeshow		
Khi-2	DDL	Pr > Khi-2
4.1746	8	0.8410

Figure 22 : Test de Hosmer Lemeshow

La figure ci-dessus montre que la p-value= 0.8410 > 0.05. Donc notre modèle est compatible avec les données.

Null deviance: 222.66 on 626 degrees of freedom
Residual deviance: 110.68 on 605 degrees of freedom

Pourcentage concordant	93.0	D de Somers	0.859
Pourcentage discordant	7.0	Gamma	0.859
Pourcentage lié	0.0	Tau-a	0.071
Paires	16200	c	0.930

Figure 23 : Concordance / Discordance (Régression logistique)

Cette figure montre que notre modèle est de très bonne qualité puisque $c = 0.93$ est supérieur à 90% et D de Somers = 0.859 est proche de 1.

1.1.2. Par la méthode forward :

Etape	Effet saisi	DDL	Nombre dans	Khi-2 du score	Pr > Khi-2	Libellé de variable
1	Capacite_Rembourseme	1	1	54.3830	<.0001	Capacite_Remboursement
2	Statut_Matrimonial	2	2	29.3242	<.0001	Statut_Matrimonial
3	Nombre_incidents_Cre	1	3	16.3874	<.0001	Nombre_incidents_Credit
4	Propriete	2	4	11.6461	0.0030	Propriete
5	Garantie	2	5	9.6130	0.0082	Garantie
6	Duree	1	6	6.1502	0.0131	Duree

Figure 24 : Récapitulatif Forward

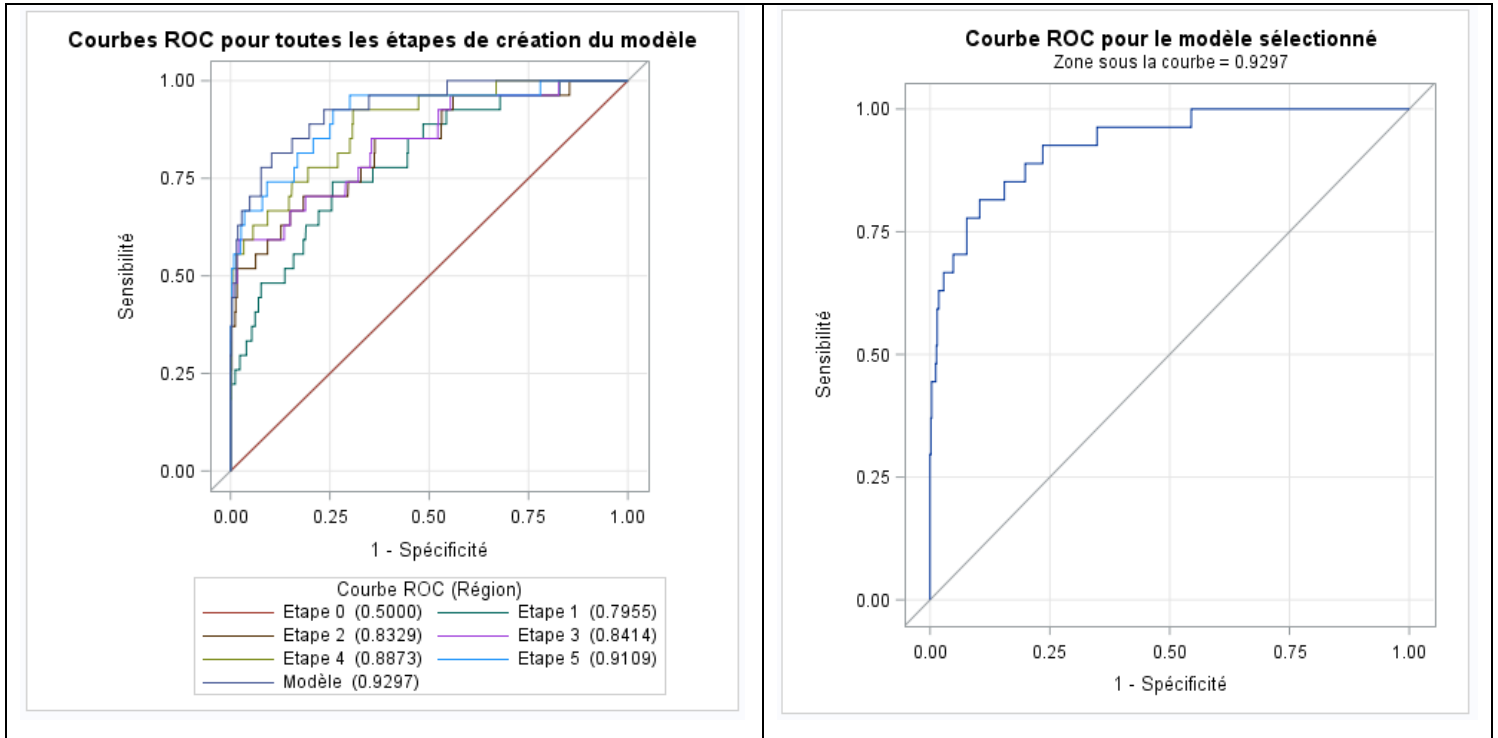


Figure 25 : Courbe ROC Forward

1.1.3. Par la méthode stepwise :

Récapitulatif sur la sélection séquentielle								
Etape	Effet		DDL	Nombre dans	Khi-2 du score	Khi-2 de Wald	Pr > Khi-2	Libellé de variable
	Saisi	Supprimé						
1	Capacite_Rembourseme		1	1	54.3830		<.0001	Capacite_Remboursement
2	Statut_Matrimonial		2	2	29.3242		<.0001	Statut_Matrimonial
3	Nombre_incidents_Cre		1	3	16.3874		<.0001	Nombre_incidents_Credit
4	Propriete		2	4	11.6461		0.0030	Propriete
5	Garantie		2	5	9.6130		0.0082	Garantie
6	Duree		1	6	6.1502		0.0131	Duree

Figure 26 : Récapitulatif Stepwise

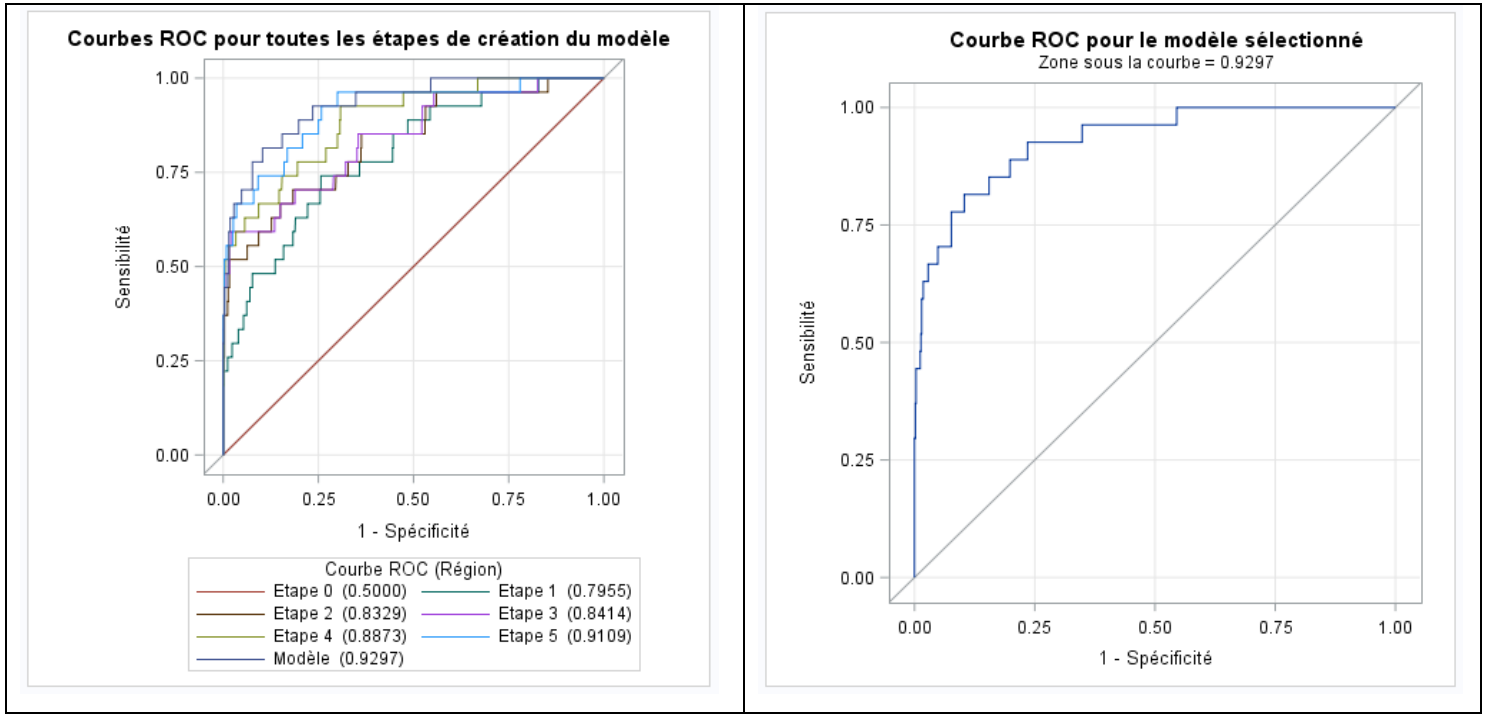


Figure 27 : Courbe ROC Stepwise

1.2. Comparaison des 3 modèles issus de la régression logistique :

Méthode de sélection	SBC	AIC	AUC	Nombre de variables
Backward	190,171	136,880	0,9297	6
Forward	184,937	140,528	0,9297	6
Stepwise	184,937	140,528	0,9297	6

Tableau 8 : Comparaison des 3 modèles sur la base de la méthode de sélection

SBC : Critère Bayésien de Schwartz (Le modèle retenu en adoptant ce critère de sélection est celui qui a le plus grand SBC)

A la suite de ces différentes régressions, nous retenons le modèle obtenu sur la base d'une sélection en arrière (Backward), car il a le plus petit AIC (et aussi le plus grand SBC). Son équation est :

$$\begin{aligned} \text{Score} = & -5,1101 - 0,00738 * \text{Durée} + 14,3563 * \text{Capacité de remboursement} \\ & + 1,0484 * \text{Nombre d'incidents sur le crédit} - 1,9381 \\ & * \text{Statut matrimonial}(\text{Marié}) - 2,4871 * \text{Propriété}(\text{Propriété}) \\ & + 2,1523 * \text{Garantie} (\text{Pas de documentation}) \end{aligned}$$

$$PD = \frac{\exp(\text{Score})}{1 + \exp(\text{Score})}$$

2. Random Forest :

Echantillonnage :

80% → Base d'apprentissage

20% → Base de test

2.1. Nombre d'arbres à considérer :

Afin de choisir le nombre d'arbres nécessaires à la construction des forêts aléatoires, nous allons visualiser la variation de l'erreur OOB en fonction du nombre d'arbres.

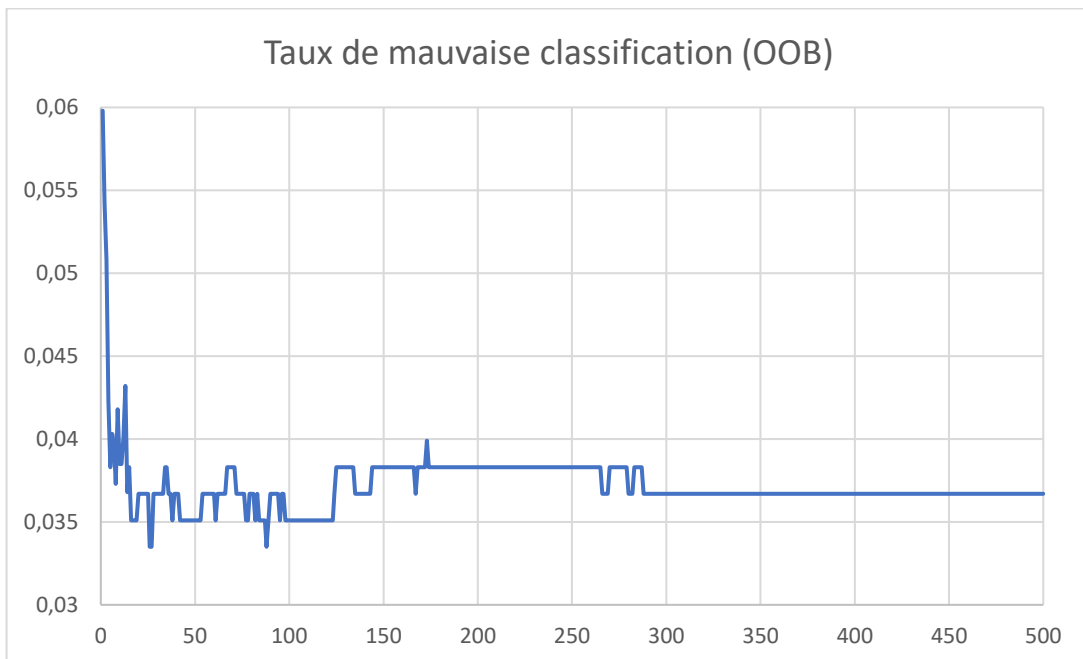


Figure 28 : Taux de mauvaises classifications en fonction du nombre d'arbres de la forêt aléatoire

Nous constatons à travers la figure ci-dessus que l'erreur se stabilise à partir de 300 arbres. Donc choisir d'appliquer la méthode de Random Forest avec 500 arbres paraît suffisant et très raisonnable.

2.2. Nombre de variables à considérer :

La méthode de Random Forest fonctionne avec un nombre limité de variables choisi de manière aléatoire pour la constitution des tests au sein de chaque nœud.

Dans le but d'obtenir le meilleur modèle c'est-à-dire celui qui minimise l'erreur (O.O.B.), nous avons construit des forêts aléatoires en faisant varier à chaque fois le nombre de variables à considérer de 1 à 16. Nous avons ainsi obtenu le résultat suivant qui schématise l'erreur o.o.b. en fonction du nombre de variables.

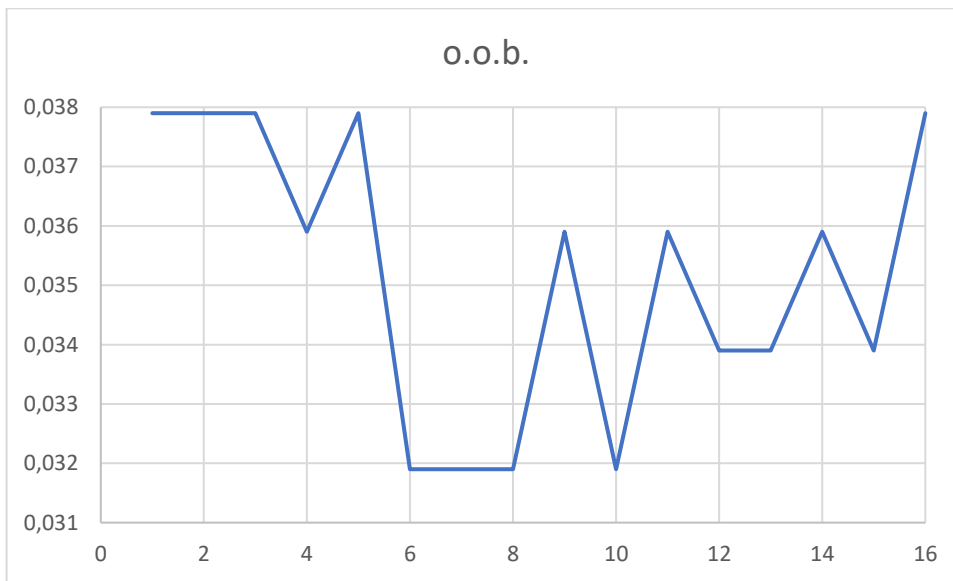


Figure 29 : Variation de l'erreur OOB en fonction du nombre de variables à considérer

Nous avons ainsi décidé de considérer 6 variables dans la construction des forêts aléatoires.

2.3. Importance des variables :

Pour mesurer l'importance de chacune des 16 variables dans la conception du modèle, évaluons le GINI associé à chacune d'elles.

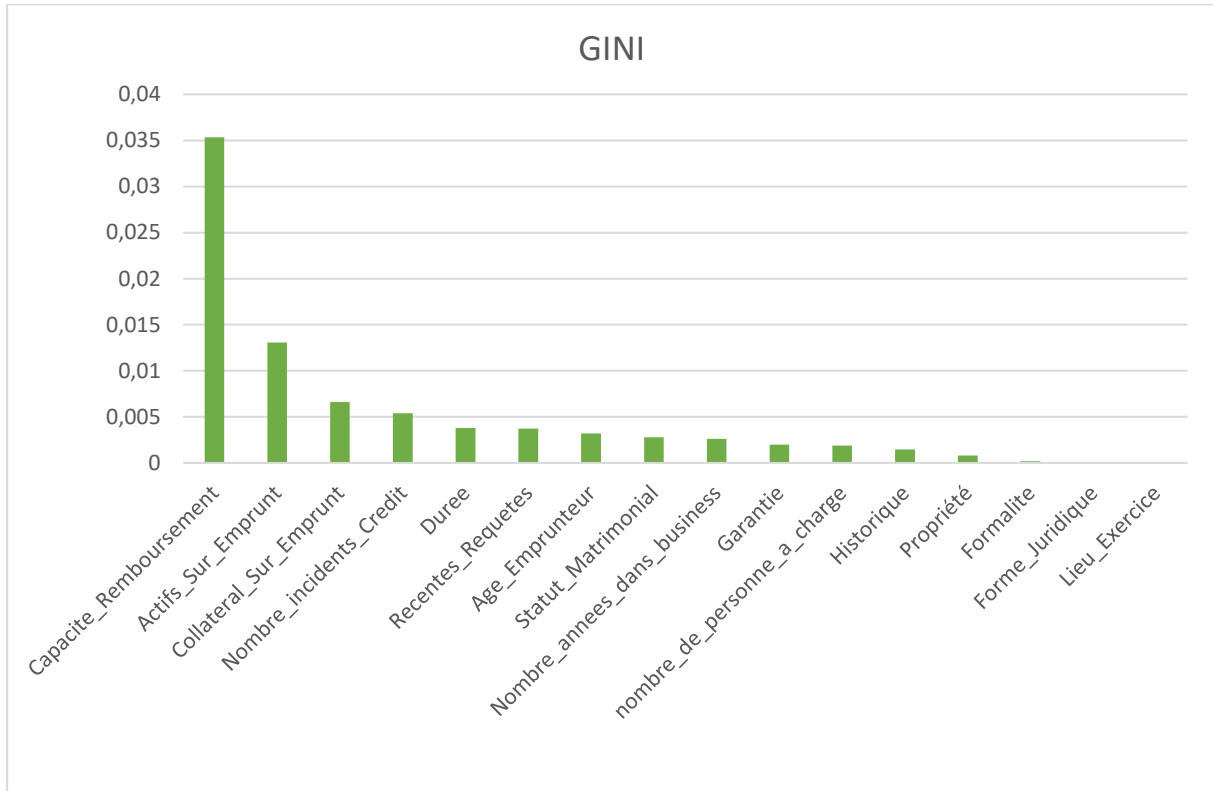


Figure 30 : Importance des variables pour la méthode RF

La variable la plus importante reste la Capacité de remboursement, en accord avec la régression logistique appliquée plus haut.

2.4. Pouvoir prédictif du modèle :

Après application de la méthode de Random Forest à notre base de données, analysons son pouvoir prédictif à travers la matrice de confusion obtenu en sortie.

Matrice de confusion :

	1	0	Class Error
1	16	3	0.185185185
0	1	481	0.001666667

$$TVP = \frac{16}{16 + 1} = 94,12\% : \text{Sensibilité}$$

$$TFP = \frac{3}{3 + 481} = 0,62\%$$

$$\text{Spécificité} = 99,38\%$$

Le taux de classification : les bons classements

$$T_c = \frac{16 + 481}{16 + 481 + 1 + 3} = 99,2\%$$

L'erreur de prédiction est de : **0,8%**

Les résultats des calculs effectués montrent qu'il s'agit d'un modèle avec un très fort pouvoir prédictif (99%).

Nous pouvons donc passer à la validation à travers la base de test (20%).

2.5. Validation sur la base de test :

En appliquant le modèle obtenu sur la base d'apprentissage ou d'élaboration du modèle sur la base de test, nous avons obtenu les résultats de prévision suivants :

Obs	Base de test	Prédiction	Obs	Base de test	Prédiction	Obs	Base de test	Prédiction
1	0	0	43	0	0	85	0	0
2	0	0	44	0	0	86	0	1
3	0	0	45	0	0	87	0	0
4	1	1	46	0	0	88	0	0
5	0	0	47	0	0	89	1	0
6	0	0	48	0	0	90	0	0
7	0	0	49	0	0	91	0	0
8	0	0	50	0	0	92	0	0
9	0	0	51	0	0	93	0	0
10	0	0	52	0	0	94	0	0
11	0	0	53	0	0	95	0	0
12	0	0	54	0	0	96	0	0
13	0	0	55	0	0	97	0	0
14	0	0	56	0	0	98	0	0
15	0	0	57	0	0	99	0	0
16	0	0	58	0	0	100	0	0
17	0	0	59	1	1	101	0	0
18	0	0	60	0	0	102	0	0
19	0	0	61	0	0	103	0	0
20	0	0	62	0	0	104	0	0
21	0	0	63	0	0	105	0	0
22	0	0	64	0	0	106	0	0
23	1	1	65	0	0	107	0	0
24	0	0	66	0	0	108	0	0
25	0	0	67	0	0	109	0	0
26	0	0	68	0	0	110	0	0
27	0	0	69	0	0	111	0	0
28	0	0	70	0	0	112	0	0
29	0	0	71	0	0	113	1	1
30	0	0	72	0	0	114	0	0
31	0	0	73	0	0	115	1	1
32	0	0	74	0	0	116	0	0
33	1	1	75	0	0	117	0	0
34	1	1	76	0	0	118	0	0
35	0	0	77	0	1	119	0	0
36	0	0	78	0	0	120	0	0
37	0	0	79	0	0	121	0	0
38	0	0	80	0	0	122	0	0
39	0	0	81	0	0	123	0	0
40	0	0	82	0	0	124	0	0
41	0	0	83	0	0	125	0	0
42	0	0	84	0	0	126	0	0

Tableau 9 : Prédiction sur base de test / Random Forest

Matrice de confusion :

	1	0	Class Error
1	7	2	0.222222222
0	1	116	0.008547009

$$TVP = \frac{7}{7 + 1} = 87,5\% : \text{Sensibilité}$$

$$TFP = \frac{2}{2 + 116} = 1,69\%$$

$$\text{Spécificité} = 98,31\%$$

Le taux de classification : les bons classements

$$T_c = \frac{7 + 116}{7 + 116 + 1 + 2} = 97,62\%$$

L'erreur de prédiction est de : **2,38%**

3. Choix du modèle final pour la grille de notation :

Modèle	Taux de bonnes prédictions
Régression logistique	93%
Random Forest	97%

Tableau 10 : Performance des 2 modèles

Il est bien vrai qu'en se référant au taux de bons classements, c'est la méthode Random Forest qui est le plus performant. Mais pour la répartition des clients en différentes catégories, il nous faut impérativement la probabilité de défaut prédite à travers le modèle. Chose qu'il n'est pas possible d'obtenir avec les forêts aléatoires.

Ainsi, pour la suite de notre travail, qui consiste à établir une grille de notation pour la classification des clients, notre modèle de référence sera celui obtenu par le biais de la régression logistique.

$$\begin{aligned} \text{Score} = & -5,1101 - 0,00738 * \text{Durée} + 14,3563 * \text{Capacité de remboursement} \\ & + 1,0484 * \text{Nombre d'incidents sur le crédit} - 1,9381 \\ & * \text{Statut matrimonial(Marié)} - 2,4871 * \text{Propriété(Propriété)} \\ & + 2,1523 * \text{Garantie (Pas de documentation)} \end{aligned}$$

$$PD = \frac{1}{1 + \exp(-\text{Score})}$$

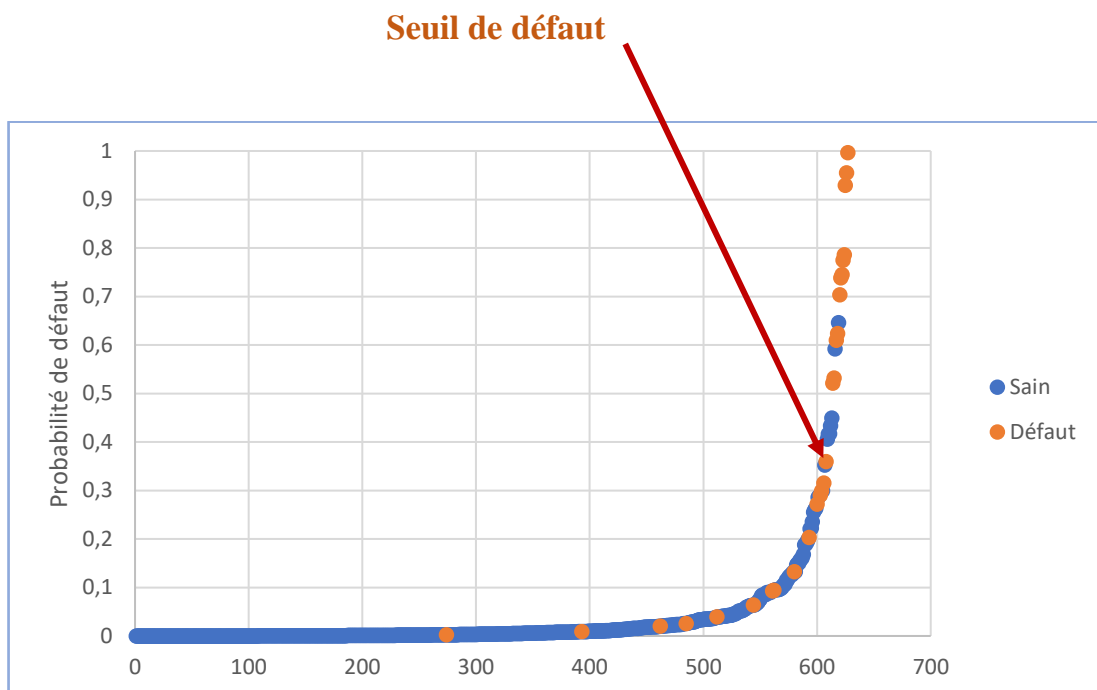


Figure 31 : Recherche du seuil de défaut

Le seuil qui réduit l'erreur de prédiction est : **0,039**

Matrice de confusion :

	1	0
1	23	93
0	4	507

$$T_c = \frac{23 + 507}{23 + 507 + 4 + 93} = \mathbf{84,52\%}$$

4. Grille de notation :

La grille de notation permet de classer les clients dans des cases spécifiques, au vu de sa probabilité de défaut, calculée par le biais du modèle choisi.

Les intervalles de PD sont choisis de telle sorte à avoir des clients sains et tous les clients distribués de façon normale en fonction de la note ; et le nombre de clients défaillants qui croient au fur et à mesure que la note baisse (de A vers D).

Pour ce faire, nous avons adopté la répartition suivante :

Note	PD	Sain	Défaut
AAA	[0 ; 0,00003]	15	0
AA]0,00003 ; 0,0001]	30	0
A]0,0001 ; 0,0003]	43	0
BBB]0,0003 ; 0,0015]	131	0
BB]0,0015 ; 0,009]	170	1
B]0,009 ; 0,03]	99	3
CCC]0,03 ; 0,1]	73	4
CC]0,1 ; 0,5]	37	7
D]0,5 ; 1]	2	12
	Total	600	27

Tableau 11 : Grille de notation

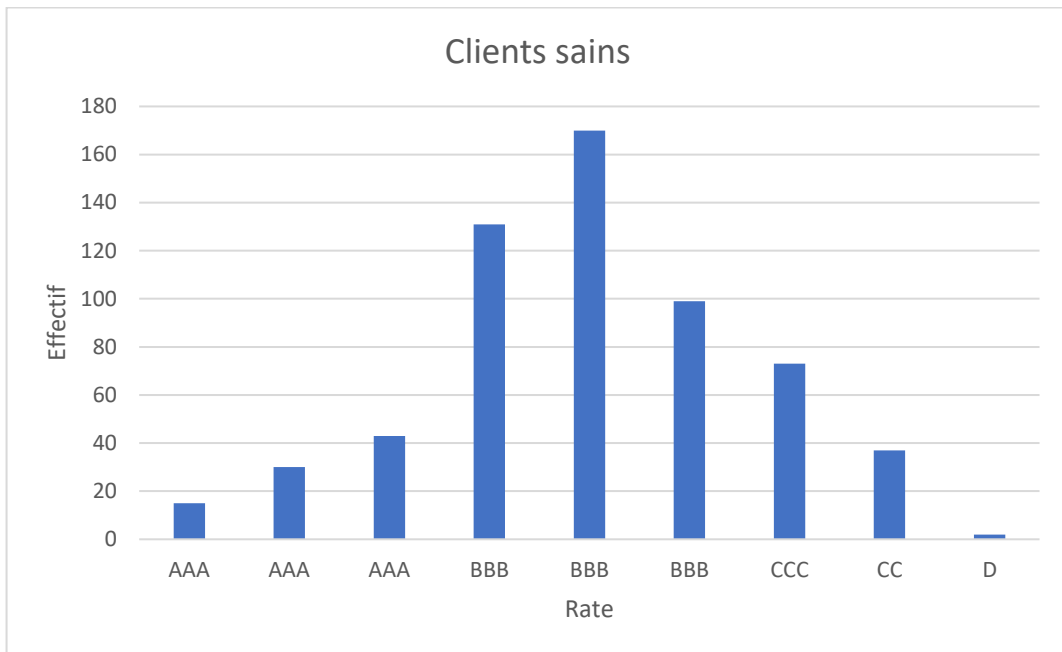


Figure 32 : Répartition des clients sains en fonction de la note

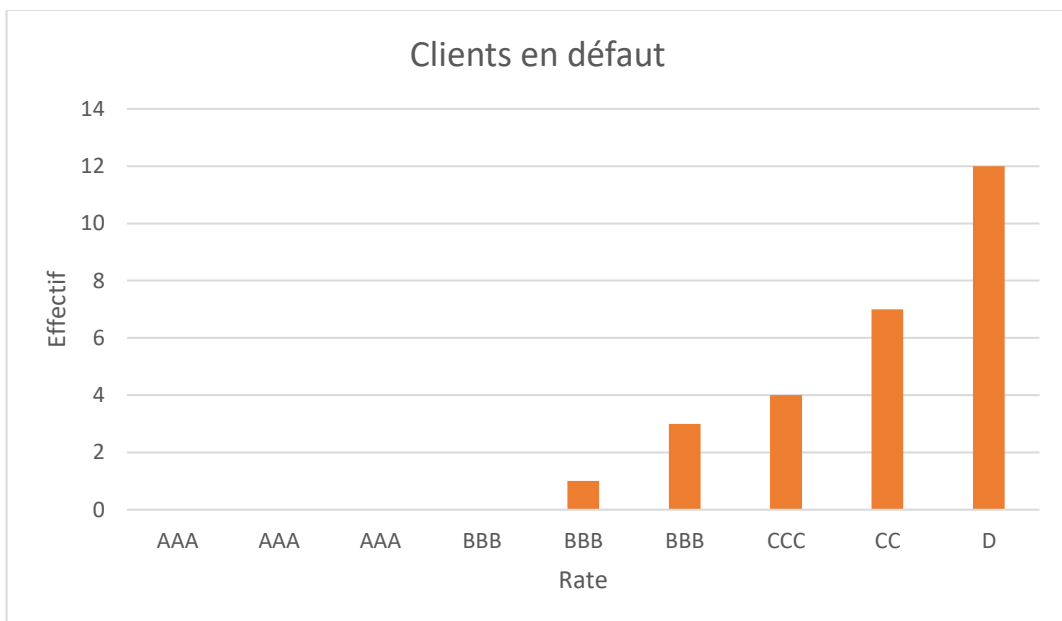


Figure 33 : Répartition des clients en défaut en fonction de la note

Décision à prendre :

Selon que le client se retrouve dans la case « verte », « jaune » ou « rouge », une décision est prise en réponse à la demande d'octroi de microcrédit.

Classe	Notes	Décision
Vert	AAA AA A BBB	Accepté
Jaune	BB B	Cas à étudier
Rouge	CCC CC D	Rejeté

Tableau 12 : Décision selon la classe d'appartenance

III. Implémentation :

Dans le but d'assurer une automatisation du système de notation et de faciliter la tâche de scoring, nous avons décidé de :

- ⊗ Concevoir une application VBA Excel. Cette dernière devra être capable de retourner en sortie la probabilité de défaut du client dont les informations seront rentrées en entrées (données relatives aux 6 variables retenues par la régression logistique) et sa note.
- ⊗ Permettre par le biais d'un bouton « Rating », de retourner la PD et la note comme dans le cas individuel, mais cette fois si, pour une base de données de plusieurs clients.

1. Notation individuelle :

L'interface se présente comme suit :

Figure 34 : Notation individuelle

En entrant les informations dans les cases aménagées pour recevoir les « INPUTS » et en cliquant ensuite sur le bouton nommé « Scoring », on obtient les résultats dans la case « OUTPUTS » (PD et Rate) comme l'indique l'image suivante.

The screenshot shows a software interface for individual scoring. At the top, the window is titled 'SCORING'. On the left, there are two input fields: 'Identifiant du client' with the value '000123456789' and 'Entreprise' with the value 'Consulting'. In the top right corner, there is a logo for 'BCP CONSULTING' featuring a horse. Below these, there are two main sections: 'INPUTS' and 'OUTPUTS'. The 'INPUTS' section, highlighted in orange, contains six fields: 'Durée de l'activité en mois' (180), 'Capacité de remboursement' (0,20861659), 'Nombre d'incidents' (0), 'Statut matrimonial' (a dropdown menu showing 'Marié'), 'Propriété' (a dropdown menu showing 'Location'), and 'Garantie' (a dropdown menu showing 'Documentation comp'). The 'OUTPUTS' section, also highlighted in orange, contains two fields: 'PD' (0,00458) and 'Rate' (BB). At the bottom right of the interface is a button labeled 'Scoring'.

Figure 35 : Sortie Scoring individuel

1. Notation par groupe de clients :

En ayant cette fois une base de données contenant, pour chaque observation, les données nécessaires au calcul de la PD, on peut noter tous les clients de la base par un simple clique sur le bouton « Rating ».

Statut_Matrimonial	Propriete	Duree	Capacite_Remboursement	Garantie	Nombre_incidents_Credit
Marié	Location	180	0,208616591	Documentation complète	0
Marié	Pas de porte	540	0,218667877	Pas de documentation	0
Marié	Propriété	108	0,033524491	Pas de documentation	0
Marié	Propriété	168	0,123101022	Pas de documentation	0
Célibataire	Location	84	0,207425743	Pas de documentation	0
Divorcé/Veuf	Location	48	0,241772459	Pas de documentation	0
Marié	Propriété	12	0,22720013	Documentation complète	0
Marié	Pas de porte	96	0,114543511	Documentation complète	0
Marié	Pas de porte	240	0,159541291	Documentation complète	0
Marié	Location	240	0,103058169	Pas de documentation	0
Marié	Location	240	0,151366686	Pas de documentation	0
Marié	Location	300	0,309226538	Documentation complète	0
Célibataire	Location	120	0,082329755	Documentation complète	0
Marié	Location	240	0,142044984	Documentation complète	0
Marié	Pas de porte	60	0,08373619	Documentation partielle	0
Marié	Location	120	0,111046386	Documentation complète	0
Marié	Pas de porte	12	0,231388564	Documentation complète	0
Marié	Pas de porte	156	0,133039981	Pas de documentation	0
Marié	Propriété	12	0,101531044	Documentation complète	0

Rating

Effacer

Figure 36 : Notation collective Scoring



Statut_Matrimonial	Propriete	Duree	Capacite_Remboursement	Garantie	Nombre_incidents_Credit	PD	Rate
Marié	Location	180	0,208616591	Documentation complète	0	0,00457922	BB
Marié	Pas de porte	540	0,218667877	Pas de documentation	0	0,00037278	BBB
Marié	Propriété	108	0,033524491	Pas de documentation	0	5,2689E-05	AA
Marié	Propriété	168	0,123101022	Pas de documentation	0	0,00012243	A
Célibataire	Location	84	0,207425743	Pas de documentation	0	0,05996513	CCC
Divorcé/Veuf	Location	48	0,241772459	Pas de documentation	0	0,53964609	D
Marié	Propriété	12	0,22720013	Documentation complète	0	0,00172275	BB
Marié	Pas de porte	96	0,114543511	Documentation complète	0	0,00221064	BB
Marié	Pas de porte	240	0,159541291	Documentation complète	0	0,00145837	BBB
Marié	Location	240	0,103058169	Pas de documentation	0	0,00555456	BB
Marié	Location	240	0,151366686	Pas de documentation	0	0,00129709	BBB
Marié	Location	300	0,309226538	Documentation complète	0	0,00797963	BB
Célibataire	Location	120	0,082329755	Documentation complète	0	0,00805201	BB
Marié	Location	240	0,142044984	Documentation complète	0	0,0011348	BBB
Marié	Pas de porte	60	0,08373619	Documentation partielle	0	0,00185344	BB
Marié	Location	120	0,111046386	Documentation complète	0	0,00176193	BB
Marié	Pas de porte	12	0,231388564	Documentation complète	0	0,02156513	B
Marié	Pas de porte	156	0,133039981	Pas de documentation	0	0,00185222	BB
Marié	Propriété	12	0,101531044	Documentation complète	0	0,000284	A

Rating

Effacer

Figure 37 : Sortie Scoring par groupe de clients

Conclusion générale

Cette étude vise à construire un modèle parcimonieux avec le plus de pouvoir prédictif pour réduire le risque lié à l'octroi de crédit. Les variables retenues dans le modèle final sont celles qui expliquent au mieux le défaut et confèrent le plus grand pouvoir prédictif à celui-ci.

Il s'agit, dans notre cas, des variables : « Capacité de remboursement », « Nombre d'incidents sur le crédit » et « Durée de l'activité en mois » comme variables quantitatives et « Statut matrimonial », « Propriété du business à financer » et « Garantie » comme variables qualitatives.

Nous avons effectué deux modélisations : une par la régression logistique et l'autre par la méthode Random Forest. Cependant, c'est le modèle issu de la régression logistique qui nous a permis d'établir la grille de notation.

Les logiciels utilisés pour ce travail sont : SAS pour la régression logistique et l'analyse exploratoire et R pour RF.

Nous avons utilisé différents tests et procédures pour valider et évaluer la précision des différents modèles obtenus. Ces tests de validation nous confortent dans le choix de la méthode appliquée. Le modèle développé au cours de cette recherche permettra d'évaluer le risque de non-remboursement pour chaque client en se basant sur son historique bancaire. Ce qui permettra à la banque d'améliorer son processus d'octroi de crédits.

Bibliographie

- [1] CREDIT RISK ANALYTICS (MEASUREMENT TECHNIQUES APPLICATIONS, and EXAMPLES in SAS), Bart BAESENS, Daniel Rosch et Harald SCHEULE, by SAS Institute, 2016
- [2] IFRS 9 and CECL Credit Risk Modelling and Validation, Tiziano Bellini, Elsevier Inc. , 2019
- [3] PRATIQUE DE LA REGRESSION LOGISTIQUE Régression Logistique Binaire et Polytomique, Ricco RAKOTOMALALA
- [4] S&P Global Rating, Annual Global Corporate Default And Rating Transition Study, 2018
- [5] Classification and Regression Trees, Leo Breiman
- [6] Random Forest, Leo Breiman , 2001
- [7] CHAOUBI, Cours GLM, INSEA MAROC, 2020

Les notations et les termes en relation

⊗ **Rating :**

Le Rating « est un mot d'origine américaine dont la traduction littérale est "évaluation". Il est défini comme un processus d'évaluation du risque attaché à un titre de créance, synthétisé en une note, permettant un classement en fonction des caractéristiques particulières du titre proposé et des garanties offertes par l'émetteur ». Il désigne à la fois un processus (l'analyse du risque) et son résultat final (la note).

⊗ **Notation financière :**

L'association française des banques retient le terme "notation" en traduction du mot rating qui veut dire « évaluation ». La notation financière, quant à elle, est une évaluation du risque de défaillance d'une émission ou d'un émetteur par une institution indépendante et spécialisée. Elle nécessite une analyse approfondie des aspects qualitatifs et quantitatifs afin d'identifier les risques de défaut. *« Les questions de base sont celles qui se posent à propos des prévisions de cash-flow que l'émetteur est susceptible de générer dans le futur, et des conditions auxquelles ces cash-flows pourront être utilisés à servir les dettes émises ».*

⊗ **Crédit scoring :**

« Le credit scoring est une méthode d'évaluation du risque de crédit. Il consiste en l'utilisation de données historiques et de techniques statistiques, dans le but d'isoler et de faire apparaître la contribution de certaines variables dans le critère de faillite ou de défaut ». Le résultat de cette application est une "fonction score" qui, génère des "scores" pour chaque emprunteur ou emprunt. Appelé-aussi probabilité de défaut, un score est un chiffre qui mesure la tendance de remboursement d'un crédit par son emprunteur. Les scores permettent le classement des emprunteurs selon la catégorie du risque, un emprunteur dont le risque de défaut est faible aura un score élevé et vice versa.

