IRM | institut du risk management

Mémoire présenté pour la validation de la Formation
« Certificat d'Expertise Actuarielle »
de l'Institut du Risk Management
et l'admission à l'Institut des actuaires
le

Par : **Benoit Piéchaud**

Titre : Modélisation de la probabilité de défaillance en assurance-crédit : une application aux secteurs de l'hôtellerie, de la restauration et du commerce de détail

Confidentialité : ☑ NON ☐ OUI (Durée : ☐ 1an ☐ 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des actuaires :

Membres présents du jury de l'Institut du Risk Management :

Secrétariat :

Bibliothèque :

Entreprise :

Nom : **Pouey International**

Signature et Cachet :

POUEY INTERNATIONAL S.A.
Société d'Exploitation
LES FILS ET PETITS-FILS DE A. POUEY
R.C. PARIS B 340 699 970
PARIS : 13, rue du Delancereaux - 75008 PARIS
BORDEAUX : 57, rue de Soissons - 33000 BORDEAUX

Directeur de mémoire en entreprise :

Nom : ***Bertrand Lacampagne***

Signature :

Invité :

Nom :

Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels
(après expiration de l'éventuel délai de confidentialité)

Signature du responsable entreprise

Signature(s) du candidat(s)

Résumé

Dans le cadre du développement par le groupe Pouey International d'une offre d'assurance-crédit spécialement adaptée aux secteurs vulnérables de l'hôtellerie, de la restauration et du commerce de détail, ce mémoire s'est attaché à la recherche d'une méthode de tarification reposant sur la probabilité de défaillance d'une entreprise.

Les différentes contraintes liées aux objectifs de l'étude et à son contexte sont la binarité de la variable à prévoir, la grande quantité de variables explicatives, l'indisponibilité fréquente des variables financières, qui ont très vite imposé d'adopter les techniques de régression logistique et des forêts aléatoires dans une optique de classification plutôt que de prévision.

Les résultats obtenus grâce à ces modèles et l'exploitation de la matrice de confusion, en particulier l'articulation entre le *taux de faux négatifs (TFN)*, le *seuil de classification* et *le taux de refus* valident la pertinence de la méthode dans le cadre de l'assurance-crédit.

En effet, appliquée à des portefeuilles de simulation réels, la méthode permet d'abaisser le taux de défaillance moyen empirique, tout en conservant un niveau d'efficience au moins aussi élevé que les analyses financières traditionnelles du groupe Pouey International.

Abstract

In the framework of the development by groupe Pouey International of a new range of trade credit insurance specifically tailored to the vulnerable sectors of hotels, restaurants and retail, this master's thesis aimed to searching for a pricing method based on the probability of default of a company.

The various constraints associated with the purpose of the study and its context: a binary dependent variable, a large quantity of explanatory variables, frequently unavailable financial data, quickly led to adopt the logistic regression and random forest methods in the perspective of classification rather than prediction.

The results achieved thanks to this models and the utilization of the confusion matrix, in particular the links between the *false negative rate (FNR),* the *classification threshold* and the *denial rate*, validate the relevance of the method in the area of trade credit insurance.

Indeed, this method applied on an existing simulation portfolio can achieve to lower the empirical average default rate, while keeping a level of efficiency at least as high as the traditional financial analysis of groupe Pouey International

Remerciements

Au moment de publier ce mémoire, je souhaite associer l'ensemble des personnes qui ont contribué directement ou indirectement à ce travail de longue haleine.

Tout d'abord, je tiens à remercier messieurs Lacampagne et Machesseau, dirigeants du groupe Pouey pour la confiance accordée et sans cesse renouvelée tout au long de ce cursus au CEA. Vous me l'avez répété, il s'agit d'un projet d'entreprise, et à ce titre j'espère que ces nouvelles compétences contribueront à l'évolution du groupe.

Mes remerciements vont également à toutes les personnes qui, dans le cadre de ce mémoire, m'ont apporté leur aide, leurs suggestions, ou leur point de vue critique et en particulier à Patrick pour son précieux appui technique.

Enfin, quelques mots pour exprimer mon affectueuse gratitude à mes proches et amis, sans qui ce projet n'aurait pu être mené à son terme. Merci pour votre soutien indéfectible et vos encouragements qui me permettent de donner le meilleur de moi-même.

Table des matières

Introduction

L'assurance-crédit est un mécanisme permettant à une entreprise de se prémunir contre le risque d'impayé portant sur son crédit client. Très développée dans certains secteurs d'activités comme la construction et l'industrie, elle reste marginale dans d'autres secteurs comme le commerce de détail, l'hôtellerie et la restauration. Ainsi sur les 700 milliards d'encours du crédit interentreprises, entre 200 et 300 milliards d'euros ne sont pas assurés. Les raisons de ce phénomène proviennent de l'attitude, à la fois des entreprises et des assureurs. D'une part, du point de vue des entreprises clientes (assurés), un certain nombre d'encours clients sont considérés comme négligeables ou en tout cas ne nécessitant pas la souscription d'une garantie. C'est en particulier le cas pour les entreprises disposant d'un portefeuille client suffisamment large avec des montants limités et pour lesquelles la solution d'auto-assurance est pertinente, la loi des grands nombres permettant de réduire la variance des pertes et donc rendant possible un provisionnement adéquat. D'autre part, du point de vue des sociétés d'assurance, certains encours sont considérés comme non assurables et font donc l'objet d'un refus. Cela est dû soit à un risque de défaillance trop élevé (et donc à une prime d'assurance qui serait considérée comme prohibitive), soit à une impossibilité technique pour l'assureur d'analyser et d'évaluer le risque de défaillance. En effet, l'octroi des garanties et leur tarification reposent sur une analyse financière et dépendent donc de la disponibilité des données financières ; or celles-ci sont souvent partiellement ou totalement indisponibles. C'est particulièrement le cas dans les secteurs du commerce de détail, de l'hôtellerie et de la restauration, qui sont principalement constitués de très petites entreprises (TPE) pour lesquelles il est très fréquent de ne pas disposer de l'ensemble des données financières (publication des comptes). En outre, en raison de leur taille et des caractéristiques de leurs activités, les encours de crédit fournisseur sont souvent de montants réduits.

Le groupe Pouey International est un acteur indépendant du marché de l'assurance-crédit dont l'implantation dans les secteurs du commerce de détail, de l'hôtellerie et de la restauration reste à développer mais qui dispose d'atouts pour cela : un prisme commercial important vers les petites entreprises et un positionnement axé principalement sur les encours clients de faibles montants.

L'objectif de ce mémoire est d'étudier la meilleure méthode de tarification de la garantie contre les impayés pour les entreprises de ces secteurs en composant avec une faible disponibilité des chiffres financiers. Il s'agit donc de modéliser la probabilité de défaillance d'une entreprise par des méthodes alternatives d'analyse crédit, c'est-à-dire ne reposant que partiellement, ou pas du tout, sur les états financiers.

Dans ce cadre, il conviendra, tout d'abord, de préciser les spécificités de l'assurance-crédit en termes de tarification et d'évaluation du risque de défaillance. La description des données financières, extra financières, endogènes et exogènes, ainsi que la définition précise de l'événement à prévoir (défaillance) permettront de se forger une première idée des différentes méthodes pertinentes.

Dès lors, différentes régressions logistiques seront testées dans une optique de prédiction ou de classification des individus (entreprises). L'analyse des statistiques de ces régressions fera clairement apparaître que des retraitements sur les échantillons d'apprentissage (biais d'antisélection) ainsi que sur les variables elles-mêmes (discrétisation, suppression des variables corrélées etc…) sont nécessaires et que les résultats sont plus pertinents à analyser sous l'angle de la classification. Par la suite une modélisation par Random Forest sera implémentée sur le même échantillon et permettra de réaliser une analyse comparative des deux méthodes. Enfin, ces différents modèles seront étudiés dans une optique de tarification d'une garantie d'assurance-crédit. Cela fera clairement apparaitre que l'analyse de la matrice de confusion est une solution indiquée puisqu'elle permet de gérer les deux leviers d'un contrat d'assurance-crédit : le taux d'acceptation et le taux de sinistralité. L'existence de différents portefeuilles de garanties déjà accordées dans les secteurs d'activité concernés permettra de réaliser une simulation et une analyse comparative de la performance des modèles retenus par rapport aux méthodes d'analyse du groupe Pouey International.

Enfin, il faudra mettre en perspective la présente étude avec les événements sanitaires survenus en2020 et leurs très probables conséquences économiques dans un futur proche. Le choix du sujet d'étude et en particulier son application aux secteurs de l'hôtellerie, de la restauration et du commerce de détail, répond à une demande interne du groupe Pouey International, bien avant que l'épisode du Covid-19 et ses impacts potentiels sur l'économie n'apparaissent. Si l'étude apparait comme brûlante d'actualité, cela est le fruit d'une coïncidence. Soulignons, en outre, que si l'intérêt principal de l'étude est la recherche d'une méthode

d'évaluation de la probabilité de défaillance dans une optique de tarification des garanties d'assurance-crédit, es résultats obtenus s'appuient sur des données dont la pertinence sera totalement remise en cause compte tenu de la probable explosion du nombre de défaillances d'entreprises. Il sera donc indispensable de réactualiser les travaux statistiques, et de s'interroger sur la robustesse du modèle dans un environnement radicalement modifié.

1. Cadre de l'étude

1.1 Présentation et fonctionnement de l'assurance-crédit

L'assurance-crédit est une branche très particulière de l'assurance non-vie qui a pour rôle de garantir le paiement des factures dans le cadre des échanges (de biens ou de services) inter-entreprises. Dans le commerce inter-entreprises (ou B to B), la facturation d'un bien ou d'un service fourni donne lieu à l'émission d'une facture, dont le paiement intervient au bout d'un délai, fixé conventionnellement par les deux parties. En France, ce délai est normalement de 45 jours maximum à partir de la date d'émission de la facture (loi LME 2005), mais de nombreuses exceptions existent dans la pratique. Dans d'autres pays, notamment en Europe du Sud (Italie, Espagne), les délais de règlement moyen sont sans commune mesure avec les délais légaux français. Dans de rares cas, un règlement est demandé au comptant, c'est-à-dire avant livraison du bien ou du service, mais il s'agit d'une pratique rare et anti-commerciale dans la plupart des écosystèmes. Il existe donc inévitablement un espace temporel durant lequel une entreprise est créancière de son client : on appelle cela le crédit inter-entreprises. Ce crédit inter-entreprises est un mode de financement non négligeable du cycle d'exploitation d'une entreprise. Au niveau macroéconomique, son importance est même supérieure au crédit bancaire : ainsi on estime que le crédit interentreprises représente environ 700 milliards d'euros alors que, dans le même temps, les encours du crédit bancaire court terme (prêts de moins d'un an, facilités de caisses, autorisations de découverts utilisées) pesaient 237 milliards d'euros à la fin février 2020 (statistiques INSEE). Tout comme le crédit bancaire, le crédit inter-entreprises comporte donc une part de risque, liée au non-remboursement des créances à échéance. En effet, entre la date d'émission de la facture et sa date d'échéance, l'entreprise débitrice peut devenir défaillante et donc en incapacité d'honorer ses dettes. Pour une banque, la gestion du risque de crédit fait partie du cœur de métier et l'analyse précédant tout octroi de crédit, ainsi que la rémunération du risque par les taux d'intérêt, lui permettent d'appréhender et de se couvrir contre ces risques. Pour une entreprise, il en va tout autrement, car d'une part cela ne constitue pas une activité clé et d'autre

part, elle ne dispose pas de fonds propres suffisants ou peut préférer les mobiliser à d'autres fins que la gestion du risque crédit. L'assurance-crédit est donc une solution permettant aux entreprises de se couvrir contre le risque de crédit commercial.

1.1.1 Principe général de l'assurance-crédit

L'assurance-crédit, contrairement aux autres branches d'assurance fait intervenir trois parties : la société d'assurance, la société assurée et la société débitrice sur laquelle porte le risque. Sa particularité résulte de ce que le risque n'est pas constitutif du client de la société d'assurance (l'assuré) mais de ses clients. Lorsqu'une entreprise souscrit une assurance-crédit, il y a potentiellement autant de risques que de clients dans le portefeuille de cette entreprise.

1.1.2 Description du marché de l'assurance-crédit

A l'origine, l'assurance-crédit a été créée dans une optique principale de gestion du risque de crédit à l'export car dans le développement d'un commerce à l'international plusieurs risques coexistent :

- le risque lié au client lui-même avec une incertitude plus grande liée à l'éloignement géographique, l'absence ou la difficulté d'obtenir des informations fiables
- le risque pays lié à sa situation économique, politique ou géopolitique et pouvant avoir des répercussions sur les entreprises et leur santé

Progressivement la demande d'assurance-crédit s'est diffusée dans le commerce intérieur, et fortement en France, ce qui s'explique sans doute par la dégradation de plusieurs facteurs notamment les crises économiques successives depuis le choc pétrolier de 1973, ayant eu pour répercussions une hausse des faillites d'entreprises et un allongement des retards de paiement.

Le secteur de l'assurance-crédit est très concentré au niveau mondial ; en effet, les trois acteurs les plus importants (Euler Hermes, Atradius et Coface) représentent à eux seuls 75% des primes collectées (*AON 2019).*

Le marché est très développé en Europe, où il est arrivé à maturité et ne connait pas d'évolution importante depuis plusieurs années Toutefois, de nombreux relais de croissance existent, notamment dans les pays émergents d'Asie et en Amérique du Sud, grâce à la diversification des canaux de distribution (l'émergence du digital, mais avant tout et surtout essor de l'affacturage). Ainsi, en 2018 les trois principaux acteurs du marché ont tous connu une augmentation du volume des primes collectées (à taux de change constant) : environ 7% pour Euler Hermès, 5% pour Atradius et 4,5% pour Coface.

En revanche, ces trois leaders sont relativement peu présents dans les deux principaux blocs économiques du XXIe siècle, les Etats-Unis et la Chine, pour des raisons différentes. Aux Etats-Unis, la pratique du crédit inter-entreprises est très limitée par des délais de paiement courts, voire par une pratique du paiement comptant généralisée. Quant à la Chine, le faible développement de l'assurance-crédit s'explique par des raisons juridiques et une politique de protectionnisme.

L'assurance-crédit est aujourd'hui une solution hybride permettant aussi bien aux grands groupes qu'aux ETI ou aux PME de se prémunir contre le risque d'impayé, pour des montants très différents et adaptés aux enjeux de chaque entreprise. Le taux de pénétration est particulièrement élevé dans les secteurs présentant la caractéristique d'une clientèle diluée et risquée, ou dans les secteurs particulièrement exportateurs. Quelques exemples :

- le commerce de gros de matériaux de construction ; le risque porte ici sur des entreprises de construction, un secteur constitué de structures fragiles, dont la statistique de défaillance est parmi les plus élevées
- le commerce de gros d'aciers dont la clientèle est constituée d'industries
- le commerce de gros de fruits et légumes dont la clientèle est constituée de revendeurs au détail
- le commerce de gros de poissons ou de viandes dont la clientèle est principalement constituée d'industries agroalimentaires

1.1.3 Tarification et gestion du risque en assurance-crédit

Lorsqu'une entreprise souscrit une assurance-crédit, elle souhaite se prémunir contre le risque d'impayé de ses clients. Pour ce faire, elle interroge la société d'assurance sur ses clients ou prospects, à la suite de quoi l'assureur réalise une analyse crédit permettant de déterminer une note de crédit, en fonction de laquelle un accord ou un refus lui sera notifié. En outre, l'analyse crédit va indiquer un montant d'exposition maximale garanti au client (sur l'entreprise analysée). En cas d'impayé, le client pourra être indemnisé jusqu'à une quote part de cette exposition maximale, en fonction de son montant créditeur envers cette entreprise à la date de l'impayé (i.e le montant des factures émises et non encore payées à la date de constatation de l'impayé).

La tarification et la gestion du risque en assurance-crédit reposent donc en grande partie sur l'analyste crédit, qui fait une partie du travail habituel des actuaires en assurance-dommages IARD classique, raison pour laquelle cette branche est souvent perçue à mi-distance entre le métier d'assurance et le métier bancaire.

La détermination de la note de risque ainsi que l'exposition garantie est le cœur de tous les modèles de tarification en assurance-crédit et peuvent aussi bien reposer sur des modèles automatisés que sur des décisions d'arbitrage à dires d'experts. En réalité, lorsque les enjeux financiers ne sont pas très importants, les sociétés d'assurance-crédit se fondent sur une analyse automatisée des éléments financiers et de leurs ratios, mais dès que l'enjeu atteint une certaine importance (tant en termes de risques qu'en termes commerciaux), l'intervention de l'humain est déterminante, comme en banque. L'intervention d'une analyste crédit, ou d'un « *arbitre* » est donc prépondérante dans un certain nombre de cas, pour trois raisons :

- les modèles actuariels automatiques peinent à tenir compte de la complexité des situations et de l'implication de variables qualitatives (type d'actionnariat, historique de l'équipe dirigeante, liens capitalistiques…).
- certains effet exogènes sont difficilement mesurables. Par exemple, il est impossible de mettre sous forme de variable le contexte politique ou géopolitique autour d'une entreprise ou d'une région. Or il peut avoir des conséquences plus radicales encore que la santé financière intrinsèque d'une entité
- l'analyse automatisée ne peut reposer que sur des données disponibles et publiées, ce qui est souvent problématique car il s'agit pour l'essentiel des états financiers publiés

une fois par an avec un retard d'au moins 6 mois. L'analyse doit donc se baser souvent sur des données anciennes

Une particularité très importante de l'assurance-crédit est que le montant du sinistre maximum pour chaque entité assurée est déterminé à l'avance. La notion de coût moyen est donc relativement simple à calculer et dépend uniquement du taux d'utilisation de l'exposition accordée par l'assureur crédit. Le principal enjeu dans le processus de tarification est donc de déterminer la probabilité de sinistre. Un modèle simple de tarification en assurance-crédit est le suivant :

$$\boldsymbol{Prime\ pure} = \boldsymbol{\pi_D} \times \boldsymbol{C} \times \boldsymbol{S}$$

Avec

- π_D la probabilité de défaillance (selon l'horizon de garantie choisi)
- C le capital (ou encours) garanti qui est fixé à l'avance et qui peut dépendre ou non de l'entreprise garantie selon le type de contrat d'assurance
- S la saturation ou le taux d'utilisation de l'encours. Ce taux est obtenu en rapportant le montant des factures impayées déclarées en cas de sinistre au capital garanti

Concernant la facturation des primes d'assurances pour les garanties contre les impayées, plusieurs méthodes cohabitent :

- une méthode basée sur le chiffre d'affaires effectif réalisé par le client avec l'entreprise couverte, sur la période de garantie (en général un an). Pour chaque exposition, une quotité du chiffre d'affaires constitue le montant de la prime. Cette quotité dépend de la sévérité du risque, mais s'établit en général autour de 0,3% du chiffre d'affaires en moyenne
- une méthode pour laquelle la facturation intervient en début de période de garantie, sur la base du montant d'exposition maximum. Elle ne dépend donc pas du chiffre d'affaires réel réalisé par l'entreprise avec son client

1.2 Présentation de l'activité du groupe Pouey International

1.2.1 Historique et activité du groupe Pouey International

Pouey Renseignement Commercial Garanti (ci-devant PRCG) est une société d'assurance créée en 1997 par la société Pouey International SA dont elle est filiale à 100%. Le groupe Pouey, fondé en 1884 est un acteur ancien du service aux entreprises, spécialisé dans la gestion du risque client. A l'origine, le groupe œuvrait principalement dans le recouvrement de créances, aussi bien dans le domaine B to B que B to C. Dans le cadre de cette activité, il réalisait de manière récurrente des enquêtes de solvabilité pour optimiser les chances de recouvrement. Finalement, ces enquêtes commerciales et financières sont devenues une prestation à part entière, commercialisée par le groupe Pouey à partir des années 1990. Puis la société a décidé d'indemniser ses clients en cas d'erreur d'appréciation. A cette fin, elle a créé en 1997 la société PRCG pour laquelle elle a obtenu un agrément d'assurance caution (branche 15). L'idée originale était d'apporter une caution aux enquêtes réalisées par sa société mère. Au départ, PRCG indemnisait d'un montant fixe prévu contractuellement le client de Pouey International en cas d'impayé à la suite d'une enquête positive. Progressivement, un montant adapté à chaque cas et selon la demande du client, a été garanti. C'est ainsi que PRCG est intervenue sur le marché de l'assurance-crédit.

Aujourd'hui, PRCG garantit des créances pour des montants allant de 3 000€ à 300 000€ en France ou à l'étranger. Le positionnement de marché de PRCG est d'intervenir en complément des assureurs crédits classiques, c'est-à-dire dans les situations de refus d'assurance, ou bien dans les cas où le montant maximum d'exposition accordé par l'assureur crédit de premier rang serait trop faible pour le client (complément).

1.2.2 Méthodes d'analyse et de tarification des garanties d'assurance-crédit du groupe Pouey International

La particularité de PRCG est d'intervenir sur des garanties refusées ou limitées par les assureurs crédits classiques. Cela a toujours pour prérequis la réalisation d'une enquête de solvabilité approfondie par Pouey International.

L'enquête est fournie au client sous la forme d'un rapport détaillé, présentant toutes les informations collectées et analysées, synthétisées en une note sur 20 et proposant un encours maximum garantissable.

L'objectif de l'enquête financière approfondie est de collecter toutes les informations pertinentes possibles concernant l'entreprise en question, afin de disposer du plus d'éléments possibles permettant une prise de décision. Les fondamentaux de l'analyse financière sont naturellement respectés : évolution de l'activité, structure de rentabilité, équilibres financiers, ratio d'endettement et de liquidité etc… Cependant, la prise en compte d'éléments qualitatifs permet bien souvent une prise de risque contre-indiquée par la simple analyse des éléments financiers. Il peut s'agir, par exemple, d'expliquer une perte annuelle par la perte d'un gros contrat, compensée par un carnet de commandes renouvelé ; ou bien encore, si l'information est recueillie, il est pertinent d'analyser quelles solutions de financement sont mises en place pour faire face à un besoin en fonds de roulement trop important (comptes courants, affacturage, découverts autorisés…).

PRCG fonde donc sa cotation de risque et, partant, la tarification de ses garanties sur une analyse humaine ou à dires d'experts. La solidité du modèle de PRCG provient de la stabilité de ses méthodes d'analyse et de cotation. Il apparait, en effet, que le taux d'erreur selon la note (« cote enquête ») est constant sur la durée. Ce taux d'erreur, déterminant la fréquence de sinistre selon la cote enquête, a permis à l'origine de bâtir le tarif d'assurance. Il a depuis très faiblement évolué, le taux d'erreur ne changeant pratiquement pas en raison de la grande stabilité des méthodes de cotation.

Si la tarification des garanties dépend principalement de la cote enquête, d'autres niveaux de segmentation ont été introduits :

- La zone géographique : France/ Europe / Grand export (reste du monde)
- La durée : 4 ou 12 mois

Concernant la facturation des primes d'assurance, c'est sur la base du montant maximum d'exposition qu'elle est établie en début de période de garantie. En aucun cas, un ajustement sur le chiffre d'affaires réellement réalisé n'est fait.

1.3 Définition de l'objectif du mémoire

1.3.1 Un produit dédié aux secteur hôtellerie, restauration et commerce de détail

La société PRCG est historiquement très ancrée dans le secteur de la construction. En effet, une grande partie de ses clients sont des entreprises de distribution de matériaux de construction, ainsi que des sociétés de travail temporaire (Intérim) dont la clientèle est risquée.

Le tableau suivant présente la part des garanties accordées secteur par secteur entre 2013 et 2018 :

Répartition des garanties par secteur d'activité	**2013** *%*	**2014** *%*	**2015** *%*	**2016** *%*	**2017** *%*	**2018** *%*
Construction	42%	44%	46%	45%	47%	49%
Industrie manufacturière	27%	25%	23%	20%	20%	19%
Commerce ; réparation d'automobiles et de motocycles	15%	14%	13%	15%	13%	12%
Activité non disponible	5%	8%	8%	11%	11%	11%
Transports et entreposage	2%	2%	3%	3%	3%	3%
Activités spécialisées, scientifiques et techniques	2%	2%	2%	2%	1%	2%
Activités de services administratifs et de soutien	2%	2%	1%	1%	2%	2%
Activités immobilières	2%	1%	1%	0%	0%	0%
Information et communication	1%	0%	0%	0%	0%	0%
Agriculture, sylviculture et pêche	1%	0%	0%	0%	0%	0%
Hébergement et restauration	0%	1%	0%	0%	0%	0%
Activités financières et d'assurance	1%	1%	0%	0%	0%	0%
Production et distribution d'eau ; assainissement, gestion des déchets et dépollution	0%	0%	0%	0%	0%	0%
Autres activités de services	0%	0%	1%	0%	0%	0%
Santé humaine et action sociale	0%	0%	0%	0%	0%	0%
Arts, spectacles et activités récréatives	0%	0%	0%	0%	0%	0%
Enseignement	0%	0%	0%	0%	0%	0%
Industries extractives	0%	0%	0%	0%	0%	0%
Production et distribution d'électricité, de gaz, de vapeur et d'air conditionné	0%	0%	0%	0%	0%	0%
Administration publique	0%	0%	0%	0%	0%	0%

On peut voir qu'il y a une prépondérance de la part du secteur de la construction dans le total des expositions. Cette part ne se réduit pas avec l'expansion commerciale de PRCG, bien au contraire elle est chaque année un peu plus importante.

Le secteur industriel est sans surprise le deuxième secteur le plus exposé dans le portefeuille de garanties de PRCG. En effet, de nombreuses sociétés de distribution d'aciers et d'autres matières premières ont recours à l'assurance de PRCG pour se couvrir contre les risques d'impayés de leurs

clients industriels, secteur particulièrement fragile et affecté par les évolutions conjoncturelles et économiques.

Le secteur du commerce de détail arrive au 3e rang des secteurs les plus exposés, mais il ne représente que 12% du total des garanties accordées en 2018. Quant au secteur de l'hôtellerie et de la restauration il est quasiment absent des garanties accordées par PRCG.

Ces deux secteurs d'activité présentent des caractéristiques assez similaires :

- Leur importance dans la population d'entreprises en France
- Une fragilité indéniable : il s'agit des deux secteurs après la construction dans lesquels le taux de défaillances est le plus élevé
- Une taille très souvent réduite aussi bien en termes de chiffre d'affaires que d'effectif
- Ils sont au bout de la chaîne économique, en contact avec le consommateur, donc étroitement lié à leur situation géographique, aux particularités régionales, à leur zone de chalandise etc…

Quel que soit le type de commerce, d'établissement d'hébergement ou de restauration, le recours à des fournisseurs de biens est indispensable et récurrent. Ces produits achetés n'étant pas payés immédiatement à la livraison au fournisseur, il existe un risque d'impayé.

Le fait que les garanties PRCG aient un taux de pénétration si faible dans ces secteurs s'explique en partie par les raisons suivantes :

- Des enjeux financiers faibles. En effet, comme il s'agit d'entreprises de petite taille, les approvisionnements n'atteignent pas des montants très importants
- Une faible concentration des encours pour le fournisseur. Par exemple, une entreprise spécialisée dans la distribution de matériel de cuisine va fournir une multitude de restaurants. Un impayé ne risque ainsi pas de mettre en danger sa pérennité
- Un coût de gestion important lié à la précision des enquêtes de solvabilité. Ce coût n'est pas adapté à une garantie de faible montant

L'enjeu pour PRCG est donc d'apporter une solution de garantie avec un coût de gestion minimal. Pour optimiser ce coût, il apparait indispensable de se passer de l'enquête approfondie et de passer à un modèle totalement automatique. De plus, il a été constaté que si l'assurance contre

les impayés pour des petits montants n'était pas vitale pour les fournisseurs de ce type de sociétés, une garantie associée à un service de relance comptable (en amont du sinistre) pouvait faire l'objet d'une réelle demande. En effet, les retards de paiement sont très courants dans ces secteurs où les entreprises sont peu structurées et où la gestion opérationnelle du quotidien prime sur la gestion administrative et financière.

L'objectif de ce mémoire est de travailler à la détermination de la meilleure méthode d'évaluation de la probabilité de défaillance qui est la principale composante du tarif (prime pure) dans le contexte présent.
Au cours de ce mémoire, la notion d'encours garanti *(« credit limit* ») ne sera pas abordée, même s'il s'agit de la deuxième composante de la prime pure.

Un des grands enjeux dans l'assurance-crédit est la disponibilité des données. En effet, la plus grande partie des analyses et donc de la détermination de la prise de risque, du montant et du tarif, repose sur les éléments financiers, issus de la publication des liasses fiscales des entreprises. Or la publication des états financiers est très souvent omise volontairement par les entreprises, dans une optique de confidentialité. En outre, la loi dite « Macron II » de 2015 est venue étendre les dérogations possibles quant à la publication des comptes annuels des sociétés. Désormais il est possible de conserver confidentiel ses comptes publiés à condition de ne pas dépasser deux des trois seuils suivants : un chiffre d'affaires de 8 millions d'euros, un total bilan de 4 millions d'euros et un effectif au cours de l'exercice de 50 personnes. Ainsi, se pose depuis toujours et encore plus sous l'empire de cette nouvelle législation, la question de l'absence de données comptables et financières dans le cadre de la modélisation de la probabilité de défaillance.

1.3.2 Définition d'un sinistre en assurance-crédit : la notion de défaillance d'entreprise

La notion de défaillance est centrale en assurance-crédit. Les contrats prévoient en général une indemnisation du montant de la créance engagée (au maximum de l'exposition garantie), dans un délai de 30, 60 ou 90 jours. S'il faut garder à l'esprit que les raisons pouvant expliquer un impayé sont multiples, deux cas peuvent se présenter :

- A la suite de l'indemnisation, les recours amiables et juridiques contre la société débitrice permettent la récupération du montant indemnisé. Le sinistre net de recours est donc de valeur nulle
- La société débitrice est en incapacité financière de régler ses engagements. Dans ce cas, une procédure collective est ouverte, ou l'a déjà été et le recouvrement des sommes indemnisées est très fortement compromis, du moins dans un délai proche

La défaillance d'entreprise étant en assurance-crédit l'aléa donnant lieu à un sinistre, il est très important de la définir précisément. Dans le cadre de ce mémoire (et de manière générale dans le cadre de l'assurance-crédit), on parlera de défaillance lorsque l'entreprise sera placée en procédure collective. Cela présente le double avantage d'être clairement défini du point de vue juridique (pas de contestation possible quant à la nature ou la survenance d'un sinistre) et temporel. Une procédure collective est une procédure juridique organisant le règlement du paiement des créances d'une entreprise lorsqu'elle se trouve dans une impasse financière (incapable de faire face à ses dettes).

Dans le domaine des procédures collectives, trois principaux statuts existent :

- La procédure de sauvegarde, lorsque l'état de cessation de paiement n'est pas encore acté mais que des difficultés financières insurmontables y conduisent inéluctablement l'entreprise. A sa propre initiative, la procédure de sauvegarde permet de poursuivre l'activité, de maintenir l'emploi et de réorganiser l'entreprise dans le cadre d'un plan arrêté par le Tribunal de Commerce
- Le redressement judiciaire, lorsque l'état de cessation de paiement est acté. Il permet, lui aussi, de sauvegarder l'entreprise en attendant que le passif soit apuré
- La liquidation judiciaire est ouverte en situation de cessation de paiement et lorsque l'activité a cessé ou lorsqu'un redressement n'est pas envisageable

Le Tribunal de commerce, qui décide de l'application de ces procédures aux entreprises, en fait la publicité par le biais des annonces légales, sur lesquelles figure la date de l'ouverture de la procédure.

1.3.3 Méthodologie et plan d'étude

1.3.3.1 Une démarche en plusieurs étapes

Dans le cadre de la recherche d'une méthode d'évaluation de la probabilité de défaillance une démarche itérative sera suivie. Dans un premier temps, des régressions logistiques seront appliquées à un même échantillon d'entreprises en y incluant, ou pas, selon les cas les différents types de variables (financières, signalétiques, exogènes…). La variable à prévoir (défaillance d'entreprise) étant binaire et les variables explicatives n'ayant pas de relation linéaire avec la défaillance, la régression logistique semble recommandée. Après avoir comparé les résultats obtenus et ainsi évalué l'apport des différents types de variables, une recherche d'amélioration sera menée en travaillant sur les variables et l'échantillon lui-même. Dans un second temps, la qualité des modèles obtenus sera évaluée en les inscrivant dans une perspective de classification et non plus d'évaluation individuelle de la probabilité de défaillance. Dans cette optique, une méthode de classification alternative (Random Forest) sera ensuite appliquée.

Enfin, la pertinence et la robustesse des modèles obtenus seront évaluées en les appliquant au contexte de PRCG et en simulant la qualité de la classification, ce qui permettra de conclure sur la bonne utilisation de ces résultats dans un contexte d'assurance-crédit.

1.3.3.2 Présentation générale des données et des échantillons mis à disposition par Pouey International

Tout d'abord, Pouey International dispose de la base de données SIRENE, fournie par l'INSEE, contenant l'ensemble des informations signalétiques de l'intégralité des entreprises de France. En effet, lors de la création d'une société, son immatriculation est obligatoire et passe par l'INSEE. Par la suite, tous les changements intervenant sur cette signalétique sont mis à jour dans la base de données SIRENE. Cette base de données est alimentée en continu par l'INSEE dans les systèmes d'information de Pouey International (nouvelles entreprises, changements…).

Concernant les défaillances, l'étude menée aura l'avantage de ne pas avoir à se limiter aux données historiques des sinistres de PRCG. En effet, les bases de données de Pouey International sont alimentées en continu par la centrale Infogreffe, qui est un groupement de tous les greffes des tribunaux de commerce de France, et qui transmettent l'ensemble des informations collectées

sur les entreprises dans le cadre de leur mission de service public. Pouey International récupère donc toutes les annonces légales ainsi que les données issues des liasses fiscales publiées (états financiers).

Fort de ces bases de données exhaustives, il n'y aura donc aucune contrainte quant à la disponibilité des données pouvant instaurer un biais. Quant à la taille des échantillons, ils seront simplement contraints par la conservation uniquement des individus pour lesquels la qualité des données est satisfaisante.

Pour rappel, ce mémoire se concentre sur deux secteurs spécifiques : le commerce de détail et l'hôtellerie et la restauration. Dans cette optique, quatre échantillons seront constitués : un échantillon d'apprentissage et un échantillon de test pour les entreprises pour lesquelles les états financiers sont disponibles, et de même les entreprises pour lesquelles ces données ne sont pas disponibles.

2. Modélisation de la probabilité de défaillance par régression logistique

2.1 Description générale de la méthode et pertinence du modèle

La régression logistique est une méthode très utilisée dans divers champs scientifiques comme l'épidémiologie, la sociologie, le marketing mais aussi la finance (gestion des risques et scoring). Elle permet de modéliser des variables binaires, c'est-à-dire suivant une loi de Bernoulli, en décrivant leur lien avec des variables explicatives. Plus précisément la régression logistique décrit les effets de chaque variable, toute chose égale par ailleurs (c'est-à-dire en tenant pour acquis les effets des autres variables explicatives). Par exemple, dans le cas de la santé, champ d'application très fréquent, la régression logistique permet l'identification des facteurs liés à une maladie ou encore de rechercher les causes de décès ou de survie de patients.

La régression logistique présente plusieurs avantages, notamment le fait que les variables explicatives ne nécessitent pas d'avoir une distribution normale, contrairement à l'analyse discriminante *(R. EISENBEIS 1977)*. De plus, la régression logistique autorise le mélange de variables qualitatives et quantitatives, ces dernières étant transformées en tranches par un processus de discrétisation.

La finalité principale de la modélisation par une régression logistique dans le cas présent est de déterminer un score permettant, d'une part, de classer les entreprises selon leur risque de défaillance, d'autre part, d'estimer une probabilité de défaillance à partir des caractéristiques de chaque entreprise.

2.2 Présentation mathématique de la régression logistique

2.2.1 La fonction LOGIT et l'estimation des paramètres

On note :

- **Y** la variable à expliquer. Celle-ci est binaire, elle peut prendre les valeurs 1 ou 0 (ou bien VRAI/FAUX, selon le domaine étudié
- **X** = $\left(X_{1,} X_{2,} \ldots . , X_{k,} \right)$ le vecteur des variables explicatives qualitatives ou quantitatives. Le modèle retient ici *k* variables explicatives.

L'objectif est ici de modéliser **π, la probabilité de défaillance, tel que** $\boldsymbol{\pi}\ (\boldsymbol{x}) = \ \boldsymbol{P}\ (\boldsymbol{Y} = \mathbf{1} \mid \boldsymbol{X} = \boldsymbol{x})$**,** où x est une réalisation du vecteur X (les valeurs prises par un certain individu).

La régression logistique propose de définir la fonction **LOGIT** de π de la manière suivante :

$\mathrm{Ln}(\frac{\pi(X)}{1-\pi(X)}) = \ a_0 + a_1 X_1 + \cdots + a_k X_k$. Ainsi $\boldsymbol{\pi}\ (\boldsymbol{x}) = \frac{\boldsymbol{e^{a_0 + a_1 X_1 + \cdots + a_k X_k}}}{\boldsymbol{1 + e^{a_0 + a_1 X_1 + \cdots + a_k X_k}}}$

L'objectif est donc d'estimer les coefficients $(a_0, a_1, \ldots, a_k)$. On peut noter que la fonction LOGIT varie de $-\infty$ à $+\infty$ donc la valeur de $\boldsymbol{\pi}\ (\boldsymbol{x})$ est bien comprise entre 0 et 1, ce qui est cohérent avec le fait qu'il s'agisse d'une probabilité. L'estimation est réalisée sur la base d'un échantillon d'apprentissage, par la méthode des maximums de vraisemblance.

Dans le contexte de l'étude, le vecteur des paramètres $A = \{a_0, a_0, \ldots, a_k\}$ est estimé par la méthode des maximums de vraisemblance dont l'expression est donnée par :

$$V(A) = \prod_{i=1}^{n} P(Y = \ y_i \mid \ X \ = x_i)$$

$$\text{soit } V(A) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left(1 - \pi(x_i)\right)^{1-y_i}$$

pour rappel, les Y_i prennent pour valeur 1 (en cas de défaillance) ou 0 (en l'absence de défaillance).

On cherche à maximiser cette fonction de vraisemblance en déterminant les coefficients qui la rendent la plus proche de 1. Cela signifie que les coefficients ainsi déterminés sont ceux qui rendent le modèle le plus proche des données d'apprentissage.

Pour simplifier la démarche, on va utiliser la log-vraisemblance, c'est-à-dire calculer le logarithme de cette fonction de vraisemblance. Maximiser la vraisemblance revient en pratique à déterminer les coefficients qui rendent la dérivée de la log-vraisemblance nulle.

$$LV(A) = \ln\big(V(A)\big) = \sum_{i=1}^{n} y_i \ln\big(\pi(x_i)\big) + (1 - y_i) \ln\big(1 - \pi(x_i)\big)$$

La résolution de ce problème est complexe et il n'existe pas de solution analytique. Dans la pratique, des méthodes d'itération existent, notamment l'algorithme de Newton-Raphson, qui permet de s'approcher par récurrence d'une solution *(R. Rakotomalala, 2017).*

2.2.2 Le choix du bon modèle en régression logistique

La qualité d'un modèle, dont la mesure est fondée sur l'analyse des écarts entre les observations et les estimations peut être déterminé par plusieurs critères distincts.

2.2.2.1 Le critère de déviance

On va comparer la qualité de l'ajustement du modèle aux données. On va d'abord considérer un modèle dit « saturé », c'est-à-dire contenant autant de variables explicatives que d'observations.

L'idée est de se baser sur la vraisemblance : plus celle-ci sera proche de 1, plus de modèle sera proche des données de l'échantillon.

Ainsi la déviance du modèle sera donnée par :

$\boldsymbol{D = 2\,(L_{sat} - L_n(\hat{A}))}$ où L_{sat} est la log-vraisemblance du modèle saturé et L_n la log-vraisemblance du modèle analysé ($\hat{A}$ est le vecteur des paramètres du modèle estimé par le maximum de vraisemblance).

Plus la déviance est faible, meilleur sera considéré le modèle en terme d'ajustement.

2.2.2.2 Le test d'Hosmer Lemeshow

On construit une statistique de test de la manière suivante :

a) Les probabilités estimées $\hat{\pi}(x_i)$ sont classées par ordre croissant
b) On sépare ensuite ces estimations en J groupe de taille égale. En général on crée des déciles, donc J = 10. On note :
    a. m_j l'effectif du groupe j
    b. o_j le nombre d'entreprises défaillante (Y = 1) observé dans ce groupe
    c. μ_j la probabilité moyenne dans le groupe j

c) La statistique de test est donnée par : $C^2 = \sum_{j=1}^{J} \frac{(o_j - m_k\mu_j)^2}{m_j\,\mu_j\,(1-\mu_j)}$

Cette statistique est enfin testée sous le Khi-deux à K-2 degré de liberté avec pour hypothèse $H_0 =$ « $le\ mod\grave{e}le\ est\ ad\acute{e}quat$ ».

Si la statistique calculée est plus grande que le quantile d'ordre 1 – α de la loi χ^2_{K-2}, alors on rejette l'hypothèse H_0.

2.2.2.3 Les critères d'Aikake et de Schwartz

Ces deux critères permettent de comparer un modèle avec un autre. Ainsi, ils sont couramment utilisés dans des processus de sélection de variables durant lesquels différents modèles sont testés en ajoutant ou supprimant des variables.

2.2.2.3.1 Critère d'information d'Akaike (AIC) :

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

Avec k : nombre de paramètres à estimer du modèle et L, le maximum de vraisemblance de ce modèle.

2.2.2.3.2 Bayesian Information Criteria (BIC) ou critère de Schwartz

$$BIC = -2\ln(L) + k\ln(n)$$

Où n est le nombre d'observation dans l'échantillon.

L'avantage de ces critères est leur faculté à optimiser le modèle, c'est-à-dire à trouver le meilleur compromis entre la précision et la complexité.

2.2.3 Procédure de sélection du meilleur modèle en régression logistique

2.2.3.1 Sélection théorique et exhaustive du meilleur modèle

Lorsqu'on dispose de ***k*** variables potentiellement explicatives dans une base de données et lorsque l'on part sans aucun parti-pris, il est normal de vouloir test l'ensemble des interactions entre chacune de ses variables dans l'optique de modéliser la variable à expliquer.

En outre, pour un modèle à ***r*** variables explicatives, on va avoir le nombre $\boldsymbol{C_r^k = \frac{k!}{r!(k-r)!}}$ correspondant au nombre de combinaison possibles et donc $\sum_{r=0}^{k} \boldsymbol{C_r^k}$, le nombre de modèles possibles.

On pourra choisir ensuite le modèle pour lequel la déviance ***D*** est maximum.

En pratique cette méthode rencontre d'importantes limites, c'est pourquoi des méthodes dites pas-à-pas sont implémentées.

2.2.3.2 Méthodes pas à pas de sélection des variables

2.2.3.2.1 Méthode descendante

Pour chaque nombre *r* de variables, au lieu de tester toutes les combinaisons possibles et de ne garder que la meilleure, on va en tester une seule.

En pratique, la méthode de sélection descendante consiste à:

- effectuer les calculs de régression pour le modèle incluant l'ensemble des variables disponibles

- réaliser un test de Student pour chacune des variables explicatives présentes. Si une ou plusieurs variables s'avèrent non significatives d'après le test, alors on enlève celle dont la statistique de Student indique qu'elle est la moins significative, puis l'on recommence le processus en enlevant à chaque étape la variable la moins significative, jusqu'au moment où l'ensemble des variables testées sont significatives. Ce dernier modèle sera celui conservé

L'avantage de cette méthode est de partir du modèle le plus complet possible et d'éliminer à chaque étape la variable la moins significative. En revanche, une variable supprimée l'est définitivement et certains modèles seraient susceptibles de fonctionner mieux avec une variable précédemment supprimée ; la non-significativité de cette variable supprimée l'étant au regard d'autres variables ayant interagit dans les modèles précédents.

2.2.3.2.2 Méthode ascendante

C'est l'exact inverse de la méthode descendante. En pratique, on va d'abord effectuer autant de régression à une seule variable, que de variables disponibles, puis on va tester le modèle le plus significatif (variable pour laquelle la statistique de Student est la plus significative). On va ensuite ajouter une variable à chaque étape, en testant chacun des modèles et ne retenant que celui où la variable supplémentaire ajoutée est la plus significative. On s'arrêtera lorsque plus aucune variable ajoutée n'est significative.

L'avantage de cette méthode est qu'elle permettra d'éviter de travailler avec plus de variables que nécessaires et que le modèle en sera amélioré à chaque étape. En revanche, une variable introduite ne peut plus être éliminée, ce qui implique le risque de passer à côté d'un modèle meilleur.

2.2.3.2.3 Procédure stpewise

L'objectif de la procédure stepwise est d'éliminer le risque de s'être séparé de variables significatives avec la méthode descendante ou la méthode ascendante. En effet, une variable peut être considérée comme très significative à une certaine étape, puis s'avérer non significative plus tard.

Ainsi, la procédure stepwise combine les caractéristiques de la méthode ascendante, c'est-à-dire l'ajout d'une variables supplémentaire (la plus significative) à chaque étape et de la méthode descendante. En effet, à chaque fois qu'on ajoute une nouvelle variable, on va tester si les variables précédemment introduites sont toujours significatives. Si des variables ne sont plus significatives, on retire du modèle la moins significative d'entre elle.

Le modèle sélectionné est celui pour lequel on ne peut plus ajouter de variables significatives, ni en retirer de non significative.

NB : on peut fonder la méthode de sélection (ascendante, descendante, stepwise) sur n'importe quel critère, par exemple AIC ou BIC. Dans ce cas, pour chaque étape du modèle, on va garder ou inversement enlever la variable pour laquelle le modèle a la mesure la plus faible ou la moins faible selon la méthode.

2.3 Travaux de traitements et d'analyse des données préalables à la régression logistique

2.3.1 Présentation des données et statistiques descriptives

2.3.1.1 Des sources de données variées

2.3.1.1.1 Les bases de données signalétiques – Répertoire SIRENE

La base de données SIRENE contient les données dites « de signalétique » de l'ensemble des entreprises et des établissements de France. L'ensemble des données sont récoltées par l'INSEE auprès des différents organismes publics et c'est également l'INSEE qui gère et maintient la base de données en la laissant accessible à l'ensemble des citoyens français (personnes physiques et morales).

Cette base de données, tout en étant disponible à l'utilisation pour le plus grand nombre, est très riche, car elle contient des informations officielles et structurées sur des millions d'établissements (11 millions actifs et 28 millions au total). De plus ces informations sont actualisées en permanence et alimentent quotidiennement les systèmes d'information de Pouey International.

Les variables communiquées via la base SIRENE sont relatives à :

- L'identification de l'entreprise (Numéro de siren, Date de création, Dénomination…)
- Des informations générales sur l'entreprise (Forme juridique, Activité Principale et secondaire, Tranche d'effectif, Catégorie entreprise…)
- La localisation géographique de l'établissement principal et des établissements secondaires
- D'autres aspects concernant les établissements.

2.3.1.1.2 Le GIE d'INFOGREFFE – publication des états financiers des entreprises

Pouey International est alimenté en permanence par son fournisseur de données Infogreffe, qui est un Groupement d'Intérêt Economique (GIE) de tous les greffes des tribunaux de commerce. Ces derniers, dans le cadre de leur mission publique, collectent un certain nombre d'informations relatives à l'ensemble des personnes morales inscrites à leur registre (RCS), qu'ils diffusent via leur GIE.

En France, il est obligatoire, sauf dispense prévue dans la loi, de déposer ses liasses fiscales annuellement auprès du greffier du tribunal de commerce où est enregistrée l'entreprise.

Ces données sont diffusées par Infogreffe sous forme de base de données, que Pouey International collecte et agrège dans ses systèmes. Pour chaque liasse fiscale publiée, ce sont environ 250 données qui sont collectées :

- Actif du bilan (*valeurs brutes – amortissements – valeurs nettes)*
- Passif du bilan
- Compte de résultat
- Détail des Immobilisations
- Détail des amortissements
- Détail des provisions
- Détail des créances et des dettes
- Détail de l'affectation du résultat & renseignements divers

2.3.1.1.3 Le BODACC et les annonces légales

Un certain nombre d'actes réalisé par les entreprises doit faire l'objet d'une publication auprès des journaux officiels. Dans beaucoup de cas, il s'agit d'une obligation légale dont la démarche est à l'initiative de l'entreprise, en particulier pour toute modification touchant aux statuts (changement de dirigeants, augmentation ou diminution du capital social, changement d'adresse du siège social, modification de la forme juridique…). Dans d'autres cas, l'annonce légale provient d'un jugement ou d'une décision prononcée par les tribunaux (redressement judiciaire, liquidation, procédure de sauvegarde).

Pouey International reçoit automatiquement et de manière quotidienne, l'ensemble des annonces légales publiées par les greffes et diffusées via le BODACC (Bulletin Officiel Des Annonces Civiles et Commerciales), avec un délai relativement court (entre 2 et 3 semaines).

2.3.1.1.4 Des données exogènes

Une autre catégorie de données peut être collectée et utilisée dans le cadre de cette étude. Elles seront qualifiées d'exogènes car elles ne se rapportent par directement aux entreprises individuellement. L'INSEE publie régulièrement des données macroéconomiques dont le niveau de granularité peut aller jusqu'à la commune. Ainsi, pour l'ensemble des communes de France le taux de chômage, le revenu moyen par habitant et l'évolution démographique de la population sur des durées de 5 et 10 ans ont pu être collectées. Il a semblé particulièrement pertinent d'inclure ce type de données dans cette étude, d'une part car il semble évident que le type d'entreprises analysées est directement impacté par la démographie, le niveau de vie et le taux d'emploi des habitants environnants (la zone de chalandise dans le contexte du commerce de détail). D'autre part, dans un contexte de raréfaction des états financiers publiés, il semblait indispensable de varier les sources de données et de ne pas se limiter aux données non financières (signalétique, annonces légales).

La plupart des données brutes n'ont que très peu de valeur prises individuellement, en revanche leur agrégation sous forme de ratio met sur un même pied d'égalité toutes les entreprises quelle que soit leur taille et ce sont ces ratios qui vont être considérés comme des variables explicatives de la défaillance, et qui vont être testées dans différents modèles.

La littérature scientifique et les travaux de recherche sur la détection des faillites sont très développés depuis assez longtemps. Un des enjeux principaux enjeux est la détermination des ratios les plus significatifs ainsi que leur pondération dans un modèle de prédiction de défaillance. C. Refait Alexandre (2004) détaille notamment les nombreuses approches qui ont été menées. Il en ressort que malgré le volume des travaux réalisés sur le sujet, aucune formule indépassable ne peut être mise en avant. En effet, le modèle adéquat dépendra de nombreux facteurs, notamment du type d'échantillon choisi ainsi que sa taille, des données disponibles, de l'horizon de prise en compte de la défaillance, mais aussi de la conjoncture économique, dont il est difficile de distinguer les déterminants. Malgré tout, certains éléments sont systématiquement mis en avant par ces travaux : la notion de rentabilité économique ou financière de l'entreprise, la structure de son bilan et sa capacité de remboursement.

2.3.1.2 Construction des bases de données

2.3.1.2.1 Description des variables explicatives

Pouey International disposant dans ses systèmes d'information de bases contenant les données de l'intégralité des entreprises (siren) enregistrées en France, actives ou non actives, le travail a donc consisté à extraire de ces bases de données toutes les entreprises des secteurs d'activité étudiés ayant publié leurs états financiers lors des 3 exercices successifs 2014, 2015 et 2016. La nécessité de disposer de 3 années de comptes publiés est un prérequis très unanimement répandu dans les théories classiques de l'analyse financière. En outre, c'est notamment les variations d'agrégats bilanciels et de certains ratios qui vont être utiles ; en particulier, la variation du chiffre d'affaires sera examinée sur 2 années successives.

Les données relatives aux annonces légales (BODACC) quant à elles sont prises depuis 2013. Elles seront considérées comme des variables qualitatives binomiales : chaque catégorie d'annonce présélectionnée est considérée comme une variable prenant la modalité 1 si l'entreprise a publié une annonce de cette catégorie depuis 2013 et 0 sinon.

Concernant les données de la base SIRENE de l'INSEE, il a été possible de récupérer la situation des données arrêtée au 31/12/2016.

Quant aux données exogènes les salaires moyens horaires par commune, ainsi que le taux de chômage sont disponibles en 2016. De même l'INSEE a publié pour les années 2006, 2011 et 2016 la démographie de chacune de ces communes. Il a donc été aisément possible de calculer les taux de variation sur 5 et 10 ans.

Pour chacune de ces entreprises, l'intégralité des données potentiellement explicatives étaient disponibles. Le « nettoyage » de l'échantillon par suppression des lignes contenant des données pas ou mal renseignées a considérablement réduit le nombre d'entreprises présentes, mais cela a tout de même été considéré comme suffisant. Dans le cadre de l'étude, trois jeux de données différents ont été constitués : un jeu ne contenant que des variables explicatives financières, c'est-à-dire les données issues des comptes publiés, un jeu ne contenant que les données non financières (signalétiques, variables macroéconomiques exogènes, annonces légales) et un jeu contenant l'ensemble des données disponibles. L'intérêt sera de tester le modèle obtenu dans des situations différentes (publication des comptes ou pas) et d'évaluer l'apport de données non financières par rapport à des travaux académiques plus classiques d'analyse financière. Dans la suite de cette étude, les trois bases de données seront appelées ainsi :

- ***Data_F*** : données financières uniquement (provenant uniquement de la publication des états financiers)
- ***Data_NF*** : données non financières uniquement (annonces légales BODACC, signalétique répertoire SIRENE, données exogènes INSEE)
- ***Data_ensemble*** : données financières et non financières

Le tableau suivant décrit le nombre d'entreprise restant dans la base à chaque étape du travail de retraitement et de préparation des jeux de données :

L'ensemble des données explicatives sont arrêtées :

LIBELLE VARIABLE	DESCRIPTION	ORIGINE	CARACTERISTIQUE
NBETABLISSEMENTS	nombre d'établissements comptant la société	INSEE - base SIRENE	variable qualitative
NatureJuridique	nature juridique (SA, SARL etc.)	INSEE - base SIRENE	variable qualitative
CAPITAL	capital social	INSEE - base SIRENE	variable continue
EFFECTIF	nombre d'ETP	INSEE - base SIRENE	variable discrète
TYPEEFFECTIF	effectif INSEE ou effectif liasses fiscales	INSEE - base SIRENE	variable qualitative
ACTIVITEPRINCIPALE	code naf principal	INSEE - base SIRENE	variable qualitative

CATEGORIEENTREPRISE	PME/ETI/Grande Entreprise	INSEE - base SIRENE	variable qualitative
INDEPENDANCEFINANCIERE	Capitaux propres / capitaux permanents avec capitaux permanents = capitaux propres + provisions pour risques et charges + dettes financières	INFOGREFFE - états financiers	variable continue
VULNERABILITEFINANCIERE	Dettes financières / Capitaux propres	INFOGREFFE - états financiers	variable continue
ENDETTEMENTCAF	Dettes financières / Capacacité d'autofinancement	INFOGREFFE - états financiers	variable continue
TAUXMARGECOMMERCIALE	Marge commerciale / Chiffre d'affaires	INFOGREFFE - états financiers	variable continue
COUVERTUREIMMOBILISATIONS	Ressources stables (Capitaux Propres + Provisions pour risques et charges + Dettes financières et comptes courants) / Actifs immobilisés	INFOGREFFE - états financiers	variable continue
PARTFINANCEMENTSTABLE	Capitaux propres + Dettes financières / Total bilan	INFOGREFFE - états financiers	variable continue
FRAISFINANCIERSCA	Charges financières (intérêts des emprunts) / Chiffres d'affaires	INFOGREFFE - états financiers	variable continue
FRAISFINANCIERSEBE	Charges financières (intérêts des emprunts) / Excédent brut d'exploitation	INFOGREFFE - états financiers	variable continue
LIQUIDITEREDUITE	Ratio = (Actif circulant - Stocks) / (Passifs à court terme)	INFOGREFFE - états financiers	variable continue
AUTONOMIEFINANCIERE	Capitaux propres/ Total bilan	INFOGREFFE - états financiers	variable continue
RESULTATEXPLOITATIONCA	Résultat d'exploitation / Chiffres d'affaires	INFOGREFFE - états financiers	variable continue
CA	Chiffres d'affaires	INFOGREFFE - états financiers	variable continue
RESULTATEXPLOITATION	Résultat d'exploitation	INFOGREFFE - états financiers	variable continue
RESULTATNET	Résultat net	INFOGREFFE - états financiers	variable continue
RESULTATEXPLOITATION	Résultat d'exploitation	INFOGREFFE - états financiers	variable continue
VALEURAJOUTEE	Valeur ajoutée	INFOGREFFE - états financiers	variable continue
PARTSALAIRESVA	Salaires et charges sociales / Valeur ajoutée	INFOGREFFE - états financiers	variable continue
EBE	Excédent brut d'exploitation	INFOGREFFE - états financiers	variable continue
CAF	Capacité d'autofinancement	INFOGREFFE - états financiers	variable continue
CAPITAUXPROPRES	Capitaux propres	INFOGREFFE - états financiers	variable continue
TOTALBILAN	Total bilan	INFOGREFFE - états financiers	variable continue
BFR	Besoin en fonds de roulement = Créances court termes - Dettes fournisseurs + Stocks	INFOGREFFE - états financiers	variable continue

BFRJOURSCA	Besoin en fonds de roulement / Chiffre d'affaires x 365	INFOGREFFE - états financiers	variable continue
TRESORERIENETTE	Trésorerie nette = Fonds de roulement - Besoin en fonds de roulement. Avec Fonds de roulement = Capitaux propres + Dettes Moyen et long terme - Actif immobilisé	INFOGREFFE - états financiers	variable continue
DELAIECOULEMENTSTOCKS	Stocks / Chiffre d'affaires x 365	INFOGREFFE - états financiers	variable continue
DELAIECOULEMENTCLIENTS	Créances clients / chiffre d'affaires x 365	INFOGREFFE - états financiers	variable continue
DELAIECOULEMENTFOURNISSEURS	Dettes fournisseurs / Chiffre d'affaires x 365	INFOGREFFE - états financiers	variable continue
RN_CA	Résultat net / Chiffre d'affaires	INFOGREFFE - états financiers	variable continue
VA_CA	Valeur ajoutée / chiffre d'affaires	INFOGREFFE - états financiers	variable continue
EBE_CA	Excédent brut d'exploitation / Chiffre d'affaires	INFOGREFFE - états financiers	variable continue
KP_TOTALBILAN	Capitaux propres / Total bilan	INFOGREFFE - états financiers	variable continue
VarCA	Variation du chiffre d'affaires entre N-1 et N	INFOGREFFE - états financiers	variable continue
VarCAN_1	Variation du chiffre d'affaires entre N-2 et N-1	INFOGREFFE - états financiers	variable continue
VarTN	Variation trésorerie nette	INFOGREFFE - états financiers	variable continue
VarKP	Variation des capitaux propres	INFOGREFFE - états financiers	variable continue
VarBFR	Variation BFR	INFOGREFFE - états financiers	variable continue
VarREX_CA	Variation du ratio résultat net / CA	INFOGREFFE - états financiers	variable continue
VarDPO	Variation du délai d'écoulement des fournisseurs	INFOGREFFE - états financiers	variable continue
VarDstock	Variation du délai d'écoulement des stocks	INFOGREFFE - états financiers	variable continue
CREATIONETS	Présence d'une Annonce légale de Création d'établissement dans la base	BODACC - annonces légales	variable qualitative
CHANGTADRESSE	Présence d'une Annonce légale de Changement d'adresse dans la base	BODACC - annonces légales	variable qualitative
CHANGTNOM	Présence d'une Annonce légale de Changement de nom dans la base	BODACC - annonces légales	variable qualitative
CHANGTACTIVITE	Présence d'une Annonce légale de Changement d'activité dans la base	BODACC - annonces légales	variable qualitative
CHANGTDIRIGEANT	Présence d'une Annonce légale de Changement de dirigeant dans la base	BODACC - annonces légales	variable qualitative

MODIFMPROD	Présence d'une Annonce légale de Modification des moyens de production dans la base	BODACC - annonces légales	variable qualitative
FERMETUREETS	Présence d'une Annonce légale de Fermeture d'établissement dans la base	BODACC - annonces légales	variable qualitative
CHANGTFJ	Présence d'une Annonce légale de Changement de la forme juridique dans la base	BODACC - annonces légales	variable qualitative
MODIFCAPITAL	Présence d'une Annonce légale de Modification de capital dans la base	BODACC - annonces légales	variable qualitative
VENTEFONDS	Présence d'une Annonce légale de Changement de Vente du fonds dans la base	BODACC - annonces légales	variable qualitative
AGE	Âge de la société	INSEE - base SIRENE	variable discrète
TxChomT42016	Taux de chômage T4 2016	INSEE - données exogènes	variable continue
EvolPop5a	Taux de variation de la population entre 2011 et 2016	INSEE - données exogènes	variable continue
EvolPop10a	Taux de variation de la population entre 2006 et 2016	INSEE - données exogènes	variable continue
SalaireMoyen	Salaire horaire moyen par habitant sur la commune	INSEE - données exogènes	variable continue

2.3.1.2.2 Définition de la variable à expliquer

Pour rappel, l'objectif de ce mémoire et de modéliser la probabilité de défaillance d'une entreprise des secteurs de l'hôtellerie, de la restauration et du commerce de détail. La notion de défaillance d'une entreprise est complexe et peut regrouper des réalités assez différentes. En effet, selon les domaines d'études ou d'application, on peut considérer qu'une entreprise est en défaillance :

- Lorsqu'elle réalise des pertes
- Lorsque ses fonds propres sont négatifs
- Lors de l'apparition du premier impayé, quel que soit le créancier concerné (banque, fournisseur, état…)
- Lors de l'ouverture d'une procédure collective

Or, quand la société d'assurance PRCG réalise une indemnisation suite à un impayé d'un client sur une société garantie, dans la grande majorité des cas, cette même société fait face à une procédure collective peu de temps après. Dans le contexte de cette étude, il convient donc de considérer l'événement à prévoir comme la survenance d'une procédure collective.

La variable à expliquer sera donc l'occurrence ou non d'une procédure collective durant l'année 2017. Pour cela, il sera fait une recherche des annonces légales suivantes dans les bases de données BODACC : *12-Liquidation judiciaire, 13-Redressement judiciaire, 16-Plan de redressement, 59- Modific procédure collective, 60-Conciliation, 61-Procédure de sauvegarde, 62-Plan de sauvegarde, 63-Liquid Judiciaire simplifiée, A3-Résol plan sauvegarde & LJ, A4-Résol plan redressement & LJ, A5- Résolution plan cession & LJ, A6- Résolution plan cession, A7-Modification plan sauvegarde, A8-Modific plan redressement, A9- Modific plan continuation, AA- Confirm REJ par Cour d'Appel, AB-Confirm LJ par Cour d'Appel, AC-Confirm SAUV par Cour d'Appel, AD-Reprise LJ apres cloture IA, AE-Extension PC au dirigeant, AF-annul plan cession retour LJ, B4-Reprise proc Liq Judiciaire, C0-Conversion en Liq Judiciaire, E2-Convers en rj de proc sauveg, E3-Faillite personnelle, E4-Faillite personnelle Loi1985, XT-resolut sauvegarde acc et LJ, XY-plan de sauvegarde acceleree, Y7-résol plan sauv fin et lj, YJ-Resol sauvegarde et REJ.*

La présence ou non d'une telle annonce sera codifiée 1 ou 0. La variable à expliquer est donc bien une variable binaire, ce qui peut autoriser sa modélisation par une régression logistique. En outre, la recherche d'une annonce de défaillance pourra être faite sur 1 ou 2 ans (seulement 2017 ou les années 2017 et 2018). Cela permettra d'élargir l'étude à la modélisation de la probabilité de défaillance sur un horizon de 2 ans.

2.3.2 Analyse primaire des échantillons et des données

2.3.2.1 Effectifs des différents échantillons en fonction des données disponibles

- **Echantillon *Data_F***

  La limitation de l'échantillon aux entreprises (Siren) dont les états financiers sont disponibles pour les années 2014, 2015 et 2016 et les retraitements associés au « nettoyage » de la base permet d'aboutir tout de même à un effectif de 46 498 individus tout de même.

Parmi cet effectif $\mathrm{N_f}$ = 46 498, il se trouve un nombre d'entreprises défaillantes lors de l'année 2017 égal à 479 soit une proportion empirique $\widehat{\mathrm{P}}_\mathrm{f}$ =1,030%.

- **Echantillon *Data_NF***

L'échantillon retenu pour les sociétés dont les informations non financières uniquement sont disponibles contient un nombre d'individus N_{nf}= 73 447. Parmi ces individus, le nombre d'annonces légales de défaillances trouvées sur les années 2017 est égal à 802 soit une proportion empirique $\widehat{\mathrm{P}}_{\mathrm{nf}} = 1{,}092\%$.

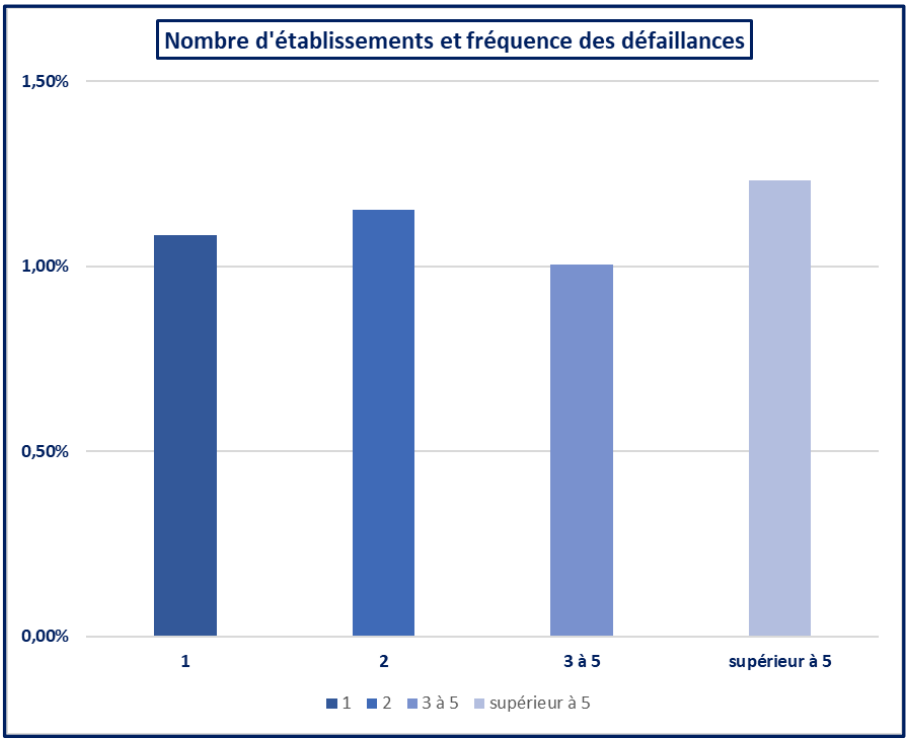
2.3.2.2 Statistiques descriptives

Dans la mesure où la variable à expliquer est qualitative binaire, la seule manière de réaliser des statistiques descriptives incluant cette variable à expliquer, est de créer des classes (ou des modalités) y compris pour les variables explicatives quantitatives, et d'observer la proportion d'entreprises défaillantes dans chacune des classes de la variable.

Par simplicité, la plupart des regroupements sont faits en fonction des quartiles.

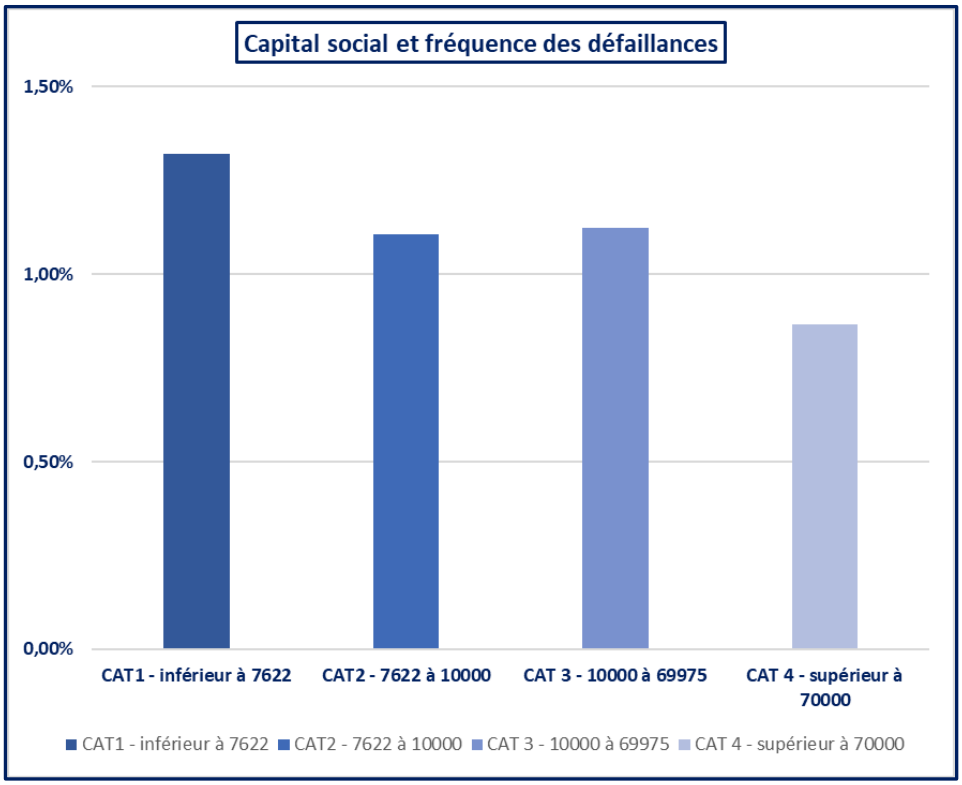
2.3.2.2.1 Statistiques sur variables non financières – Echantillon *Data_NF*

2.3.2.2.1.1 Nombre d'établissements déclarés

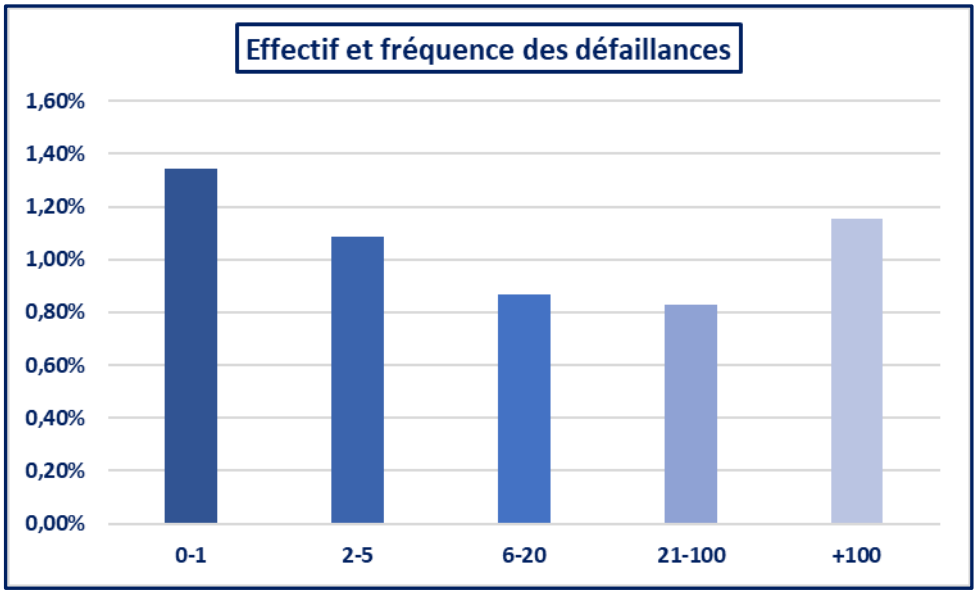


Il ne semble pas y avoir d'incidence visible du nombre d'établissements sur la fréquence des défaillances constatées sur 1 an.

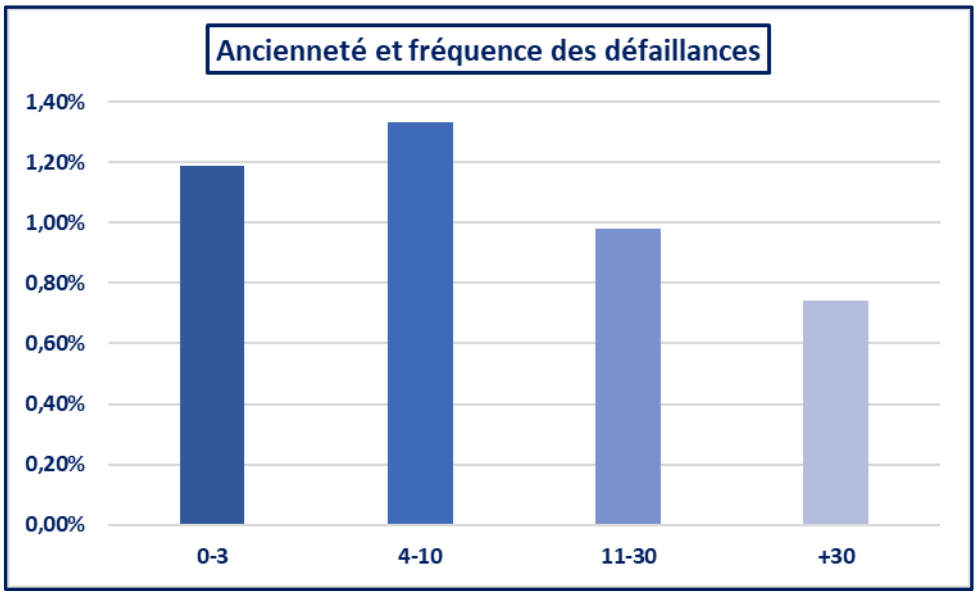
2.3.2.2.1.2 Montant du capital social



2.3.2.2.1.3 Effectif de l'entreprise

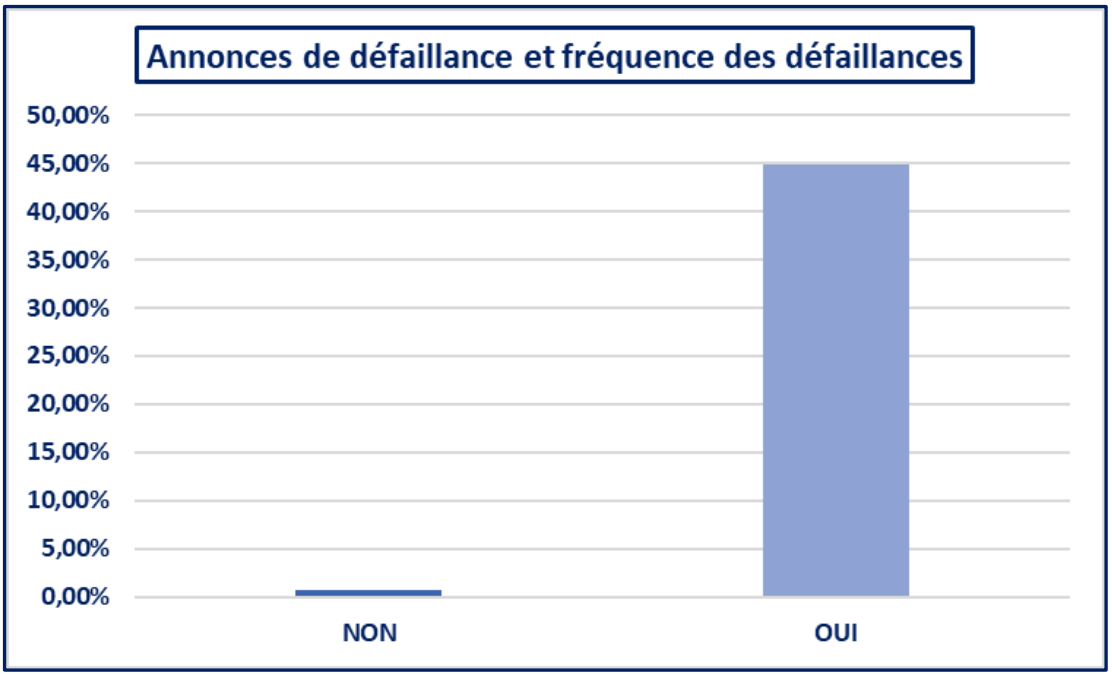


2.3.2.2.1.4 Ancienneté de l'entreprise



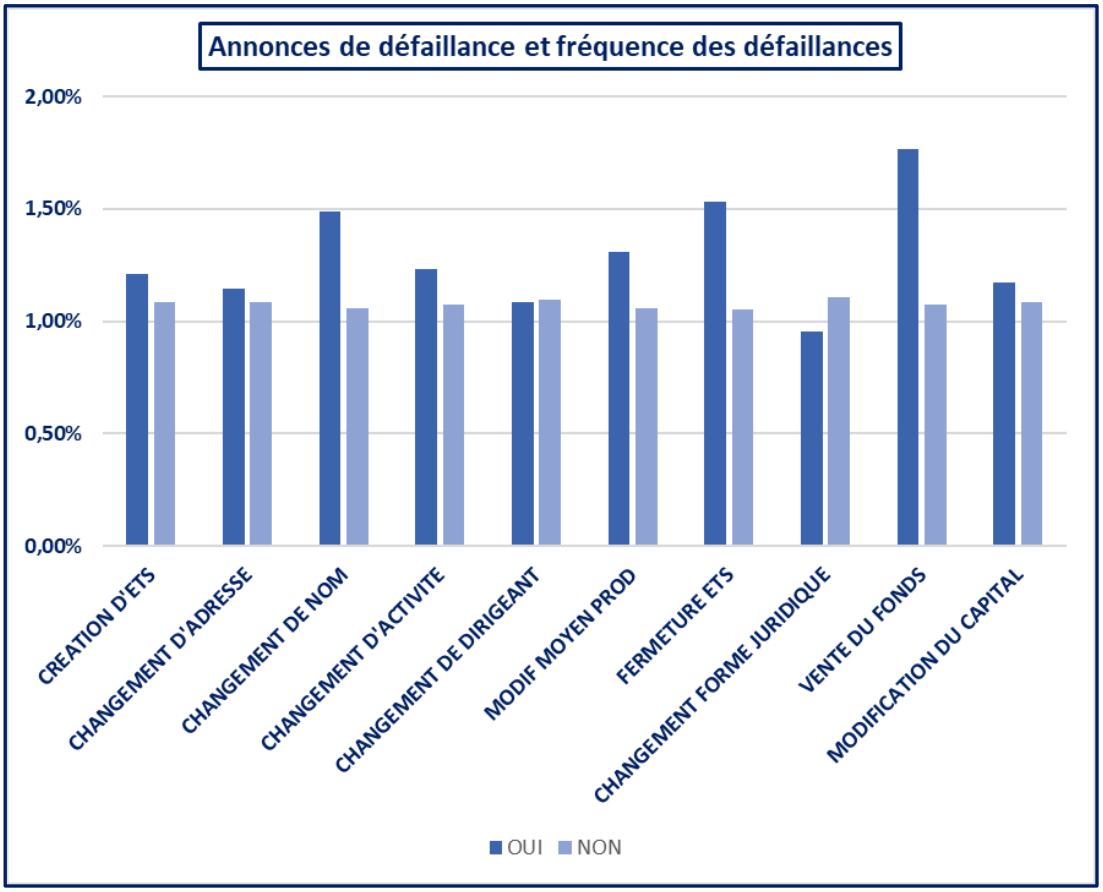
Il y a visiblement une relation quasi linéaire entre le taux de défaillance et l'ancienneté de l'entreprise. En effet, les entreprises les plus anciennes (au moins 30 ans d'ancienneté) ont un taux de défaillance beaucoup moins important que les plus récentes (de 0 à 10 ans d'ancienneté).

2.3.2.2.1.5 Annonce de défaillance dans les 4 dernières années



Sur les 759 entreprises de l'échantillon ayant rencontré une défaillance antérieurement (depuis 2013), mais toujours en activité, présentes dans l'échantillon, 340 sont défaillantes l'année suivante. C'est un constat peu étonnant dans la mesure où la plupart des défaillances d'entreprises se font en deux étapes et sont donc marquées par deux annonces : Plan de sauvegarde ou Redressement judiciaire, puis Liquidation Judicaire.

2.3.2.2.1.6 Annonces légales diverses



La présence d'annonce légales diverses ne semble pas avoir d'influence très significative sur le taux de défaillance d'entreprises dans l'échantillon étudié, même s'il peut y avoir un taux plus élevé :

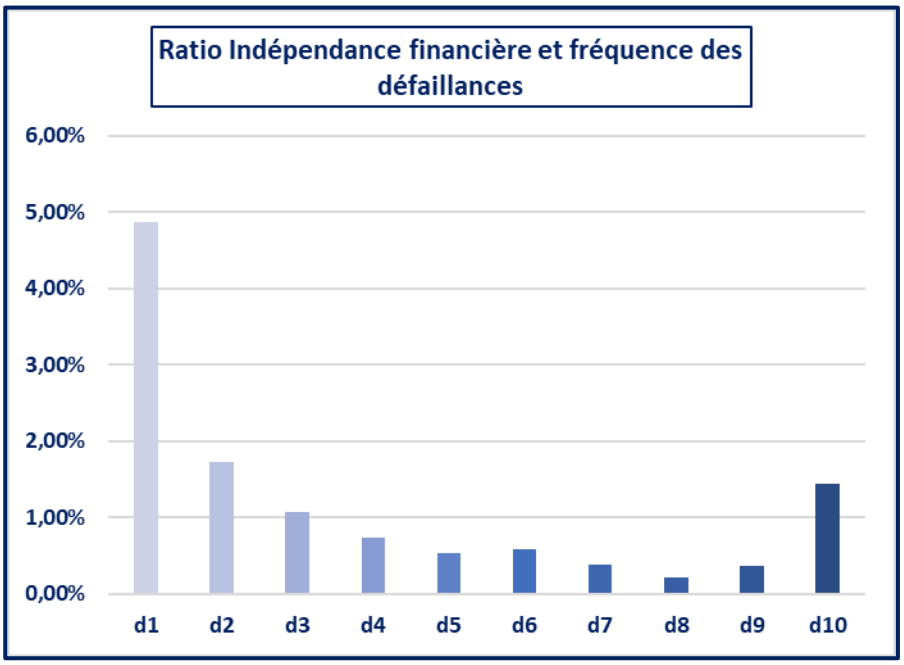
- pour les entreprises ayant connu un changement de propriétaire du fonds (vente du fonds)
- pour les entreprises ayant fermé un établissement
- pour les entreprises ayant changé de nom

2.3.2.2.2 Statistiques descriptives sur variables financières – Echantillon *Data_F*

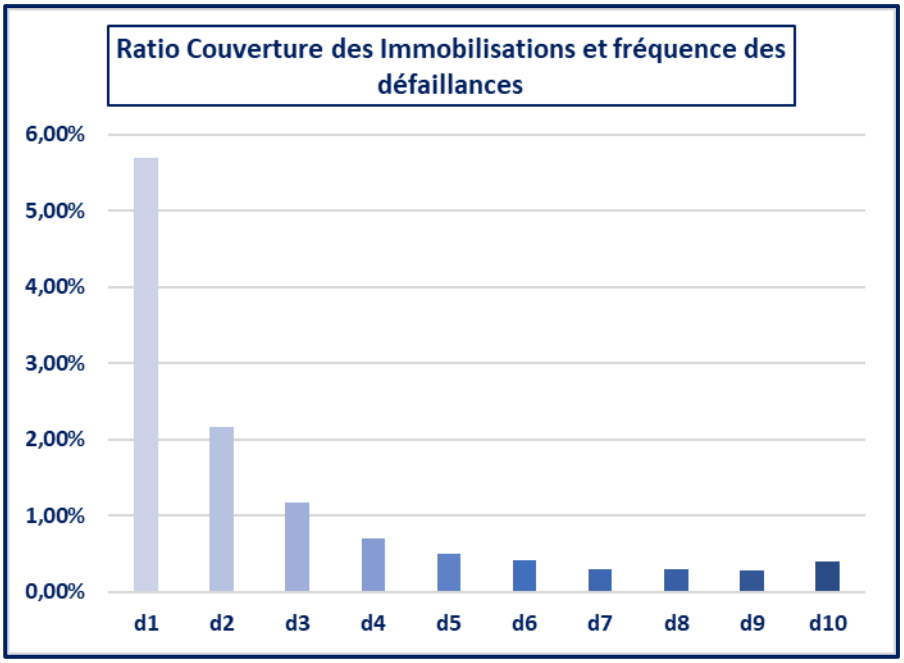
Pour obtenir une première analyse de l'impact d'une variable dans la probabilité de défaillance ces variables sont découpées en déciles. En effet, les variables issues des états financiers sont évidemment toutes quantitatives, mais dans la mesure où la variable à expliquer est binaire (DEFAILLANCE OU NON DEFAILLANCE), il est nécessaire de discrétiser ces variables et donc de créer des catégories au sein desquelles le taux de défaillance moyen sera observé. L'échantillon contenant 56 641 SIREN, un décile en contiendra donc 5664, ce qui est suffisant pour des observations pertinentes.

Dans la modélisation par régression logistique, l'ensemble des variables financières sont inclues mais ici le choix est fait de ne représenter que celle ayant le comportement le plus significatif.

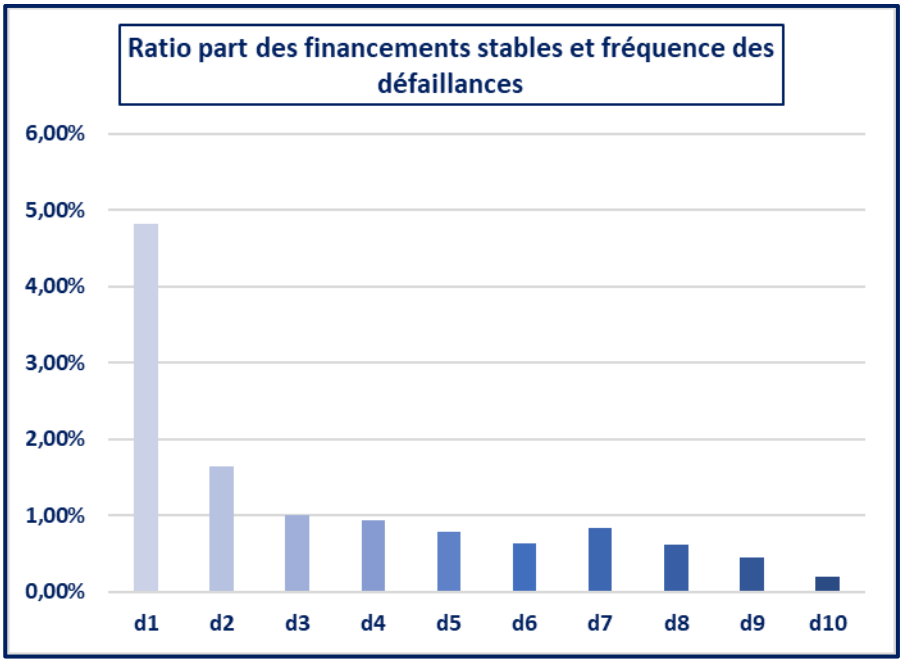
2.3.2.2.2.1 Ratio d'indépendance financière



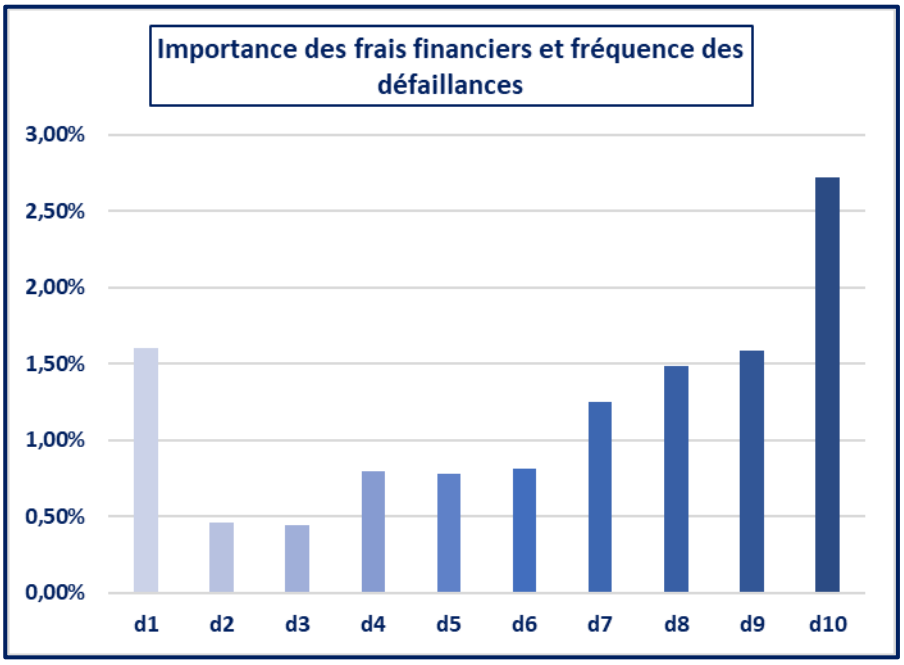
2.3.2.2.2.2 Couverture des immobilisations



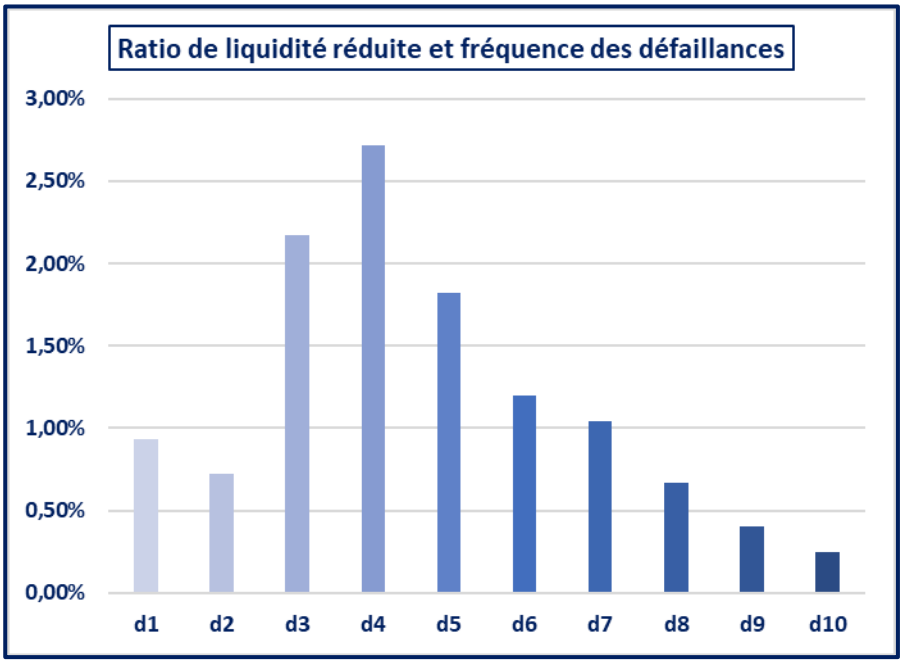
2.3.2.2.2.3 Part des financements stables



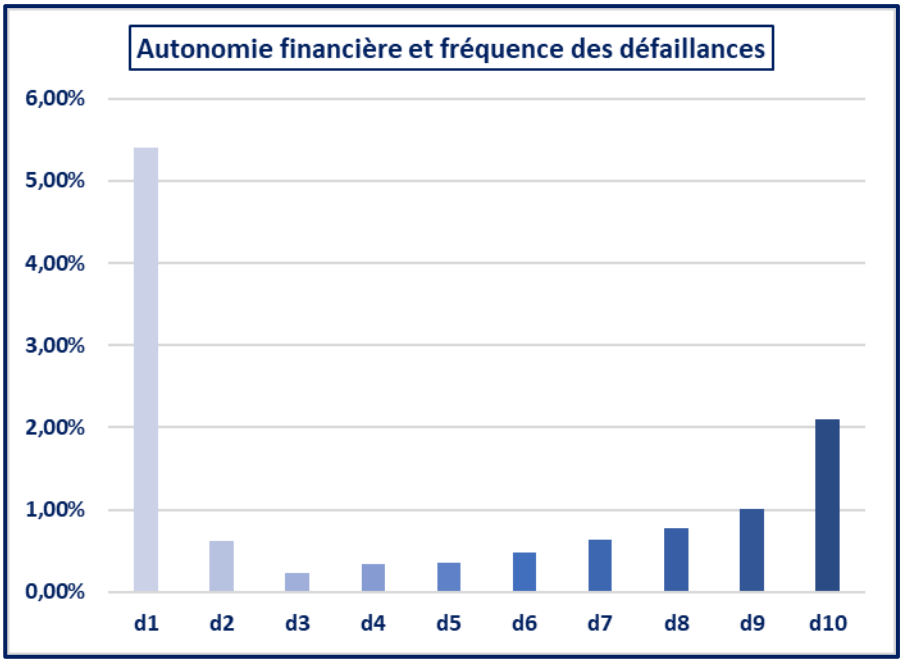
2.3.2.2.2.4 Importance des frais financiers



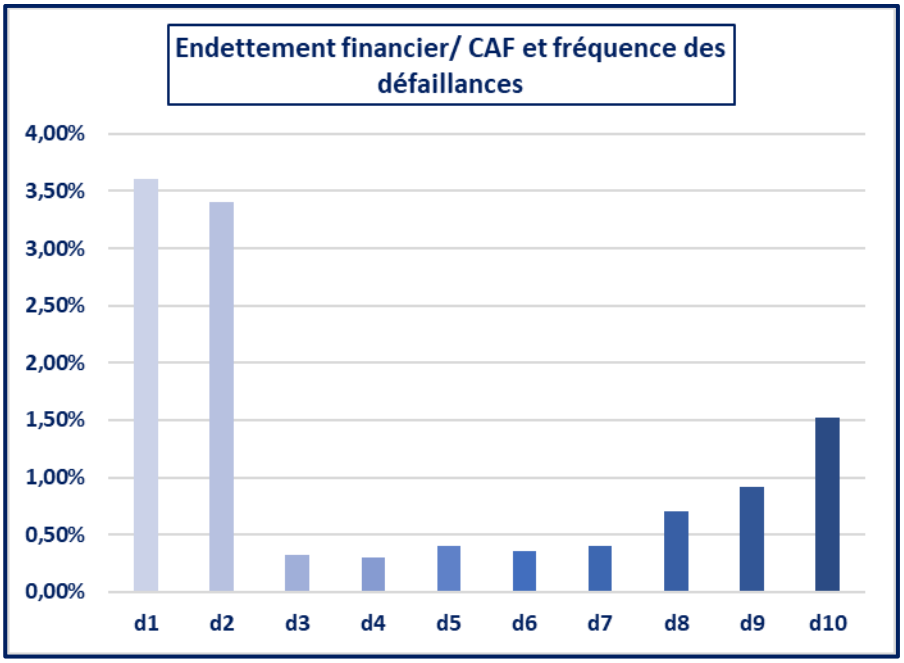
2.3.2.2.2.5 Ratio de liquidité réduite



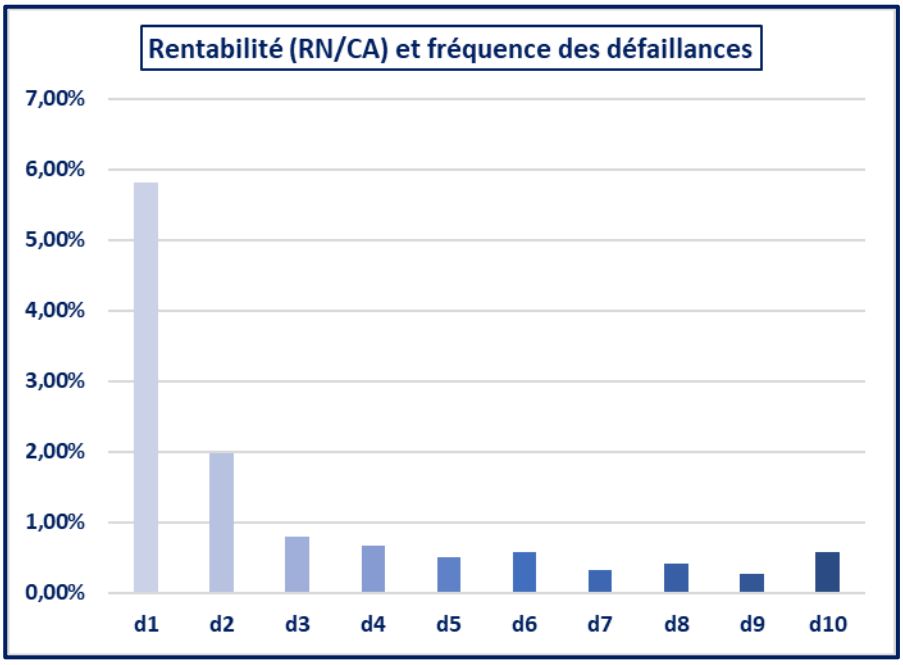
2.3.2.2.2.6 Autonomie financière



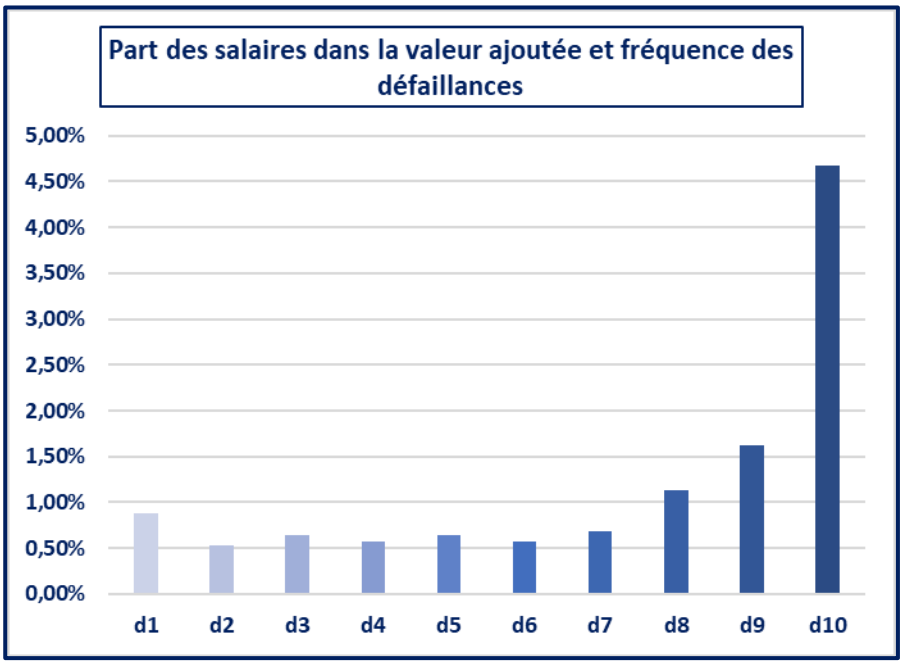
2.3.2.2.2.7 Ratio d'endettement sur Capacité d'autofinancement (CAF)



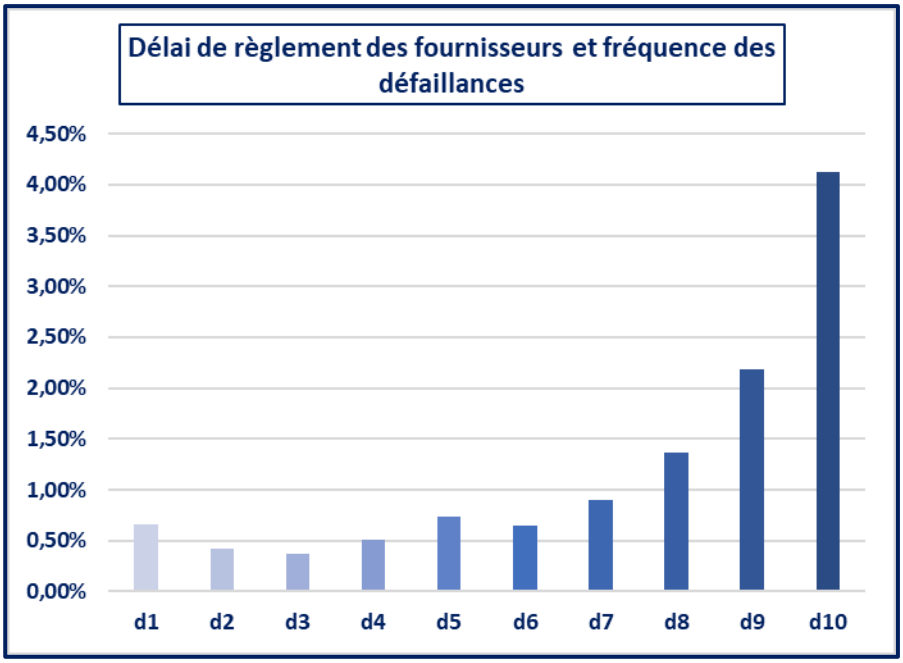
2.3.2.2.2.8 Résultat net / Chiffre d'Affaires



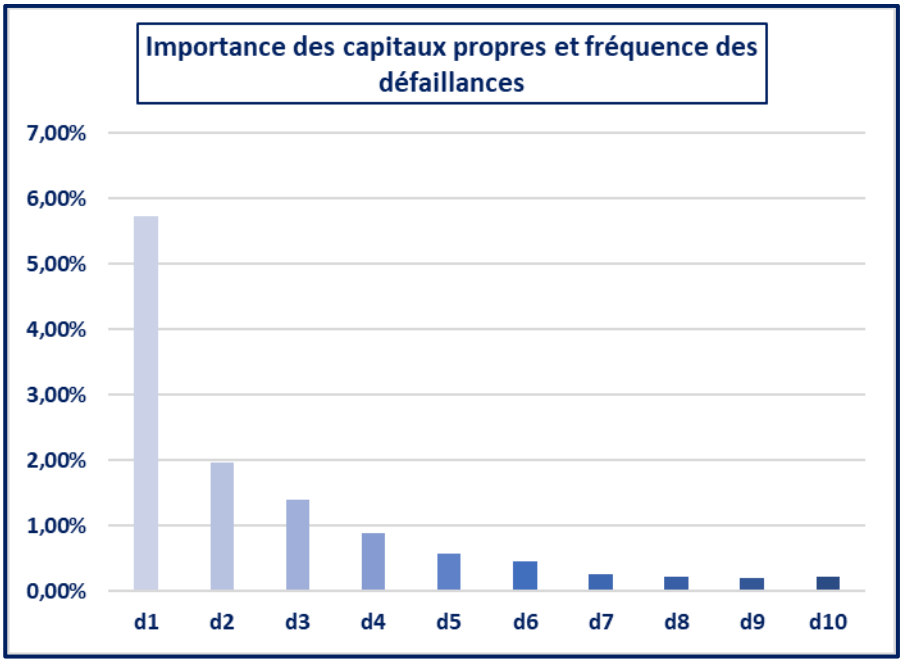
2.3.2.2.2.9 Part des salaires dans la VA



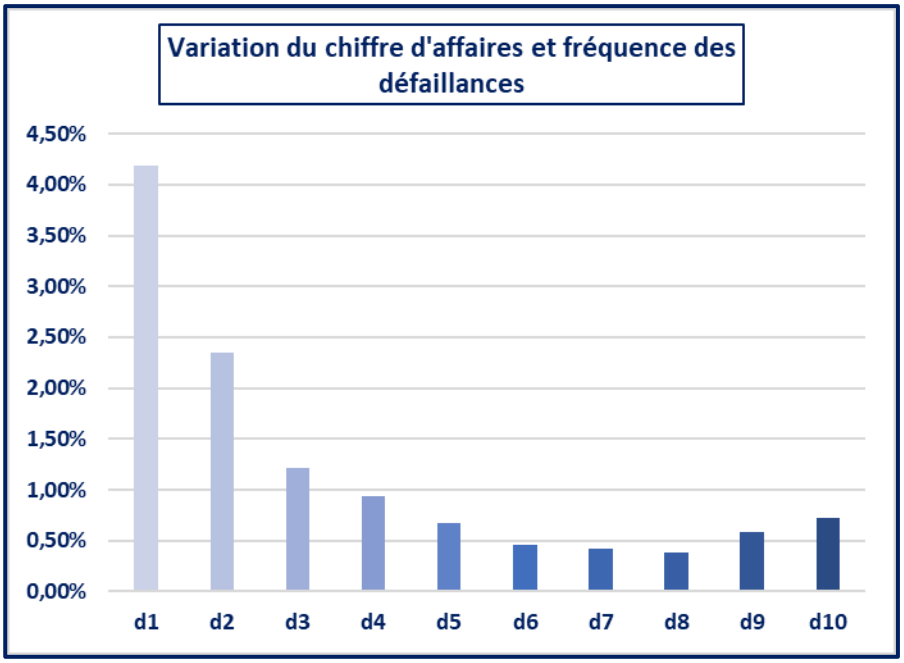
2.3.2.2.2.10Délai d'écoulement des fournisseurs



2.3.2.2.2.11Capitaux propres / Total bilan



2.3.2.2.2.12Variation du chiffre d'affaires



2.3.2.3 Analyse bi-variée

L'objectif de cette démarche est de déceler les variables fortement corrélées. Lorsque deux variables sont fortement corrélées, l'inclusion de ces deux variables dans un modèle GLM n'apporte pas réellement d'informations : il est donc possible de n'en conserver qu'une. Dans le contexte présent, cela permettra de réduire le nombre de variable élevé.

L'étude de la corrélation des différentes variables est réalisée avec le logiciel R, et, dans la mesure où une partie des variables sont de type ordinal, c'est le coefficient de corrélation de Spearman qui est utilisé ; en effet, celui-ci rend mieux compte de l'existence d'une relation non linéaire que le coefficient de Pearson.

Principe :

Considérons deux variables explicatives X et Y. Pour l'ensemble des valeurs X_i et Y_i prises pour les n individus de la série ($i \in [1; n]$), un rang $R(X_i), R(Y_i)$ va être assigné ce qui donnera une série bi-variée $\{R(X_i), R(Y_i); i = 1, \dots, n\}$.

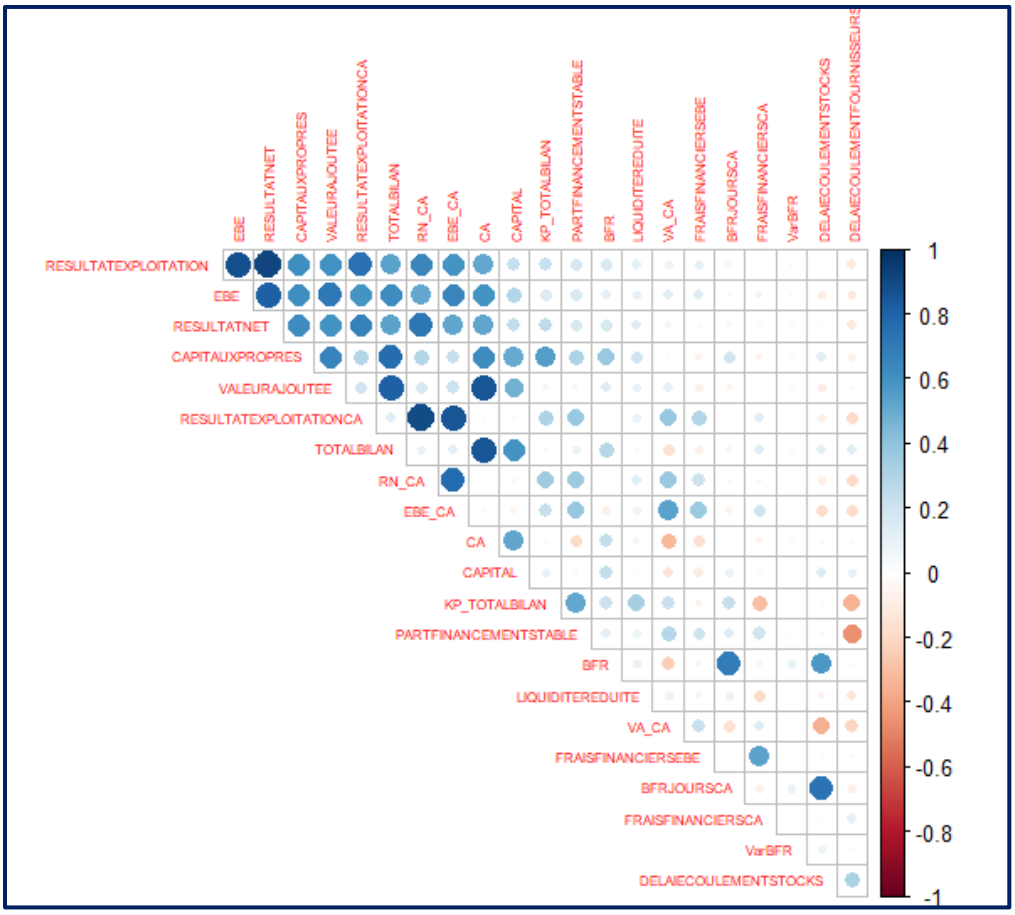
On définit ainsi le coefficient de Spearman :

$$r_s(X,Y) = \frac{\frac{1}{n}\sum_i \left(R(X_i) - \overline{R_X}\right) \times \left(\left(R(Y_i) - \overline{R_Y}\right)\right.}{\sqrt{\frac{1}{n}\sum_i (R(X_i) - \overline{R_X})^2} \times \sqrt{\frac{1}{n}\sum_i (R(Y_i) - \overline{Y})^2}}$$

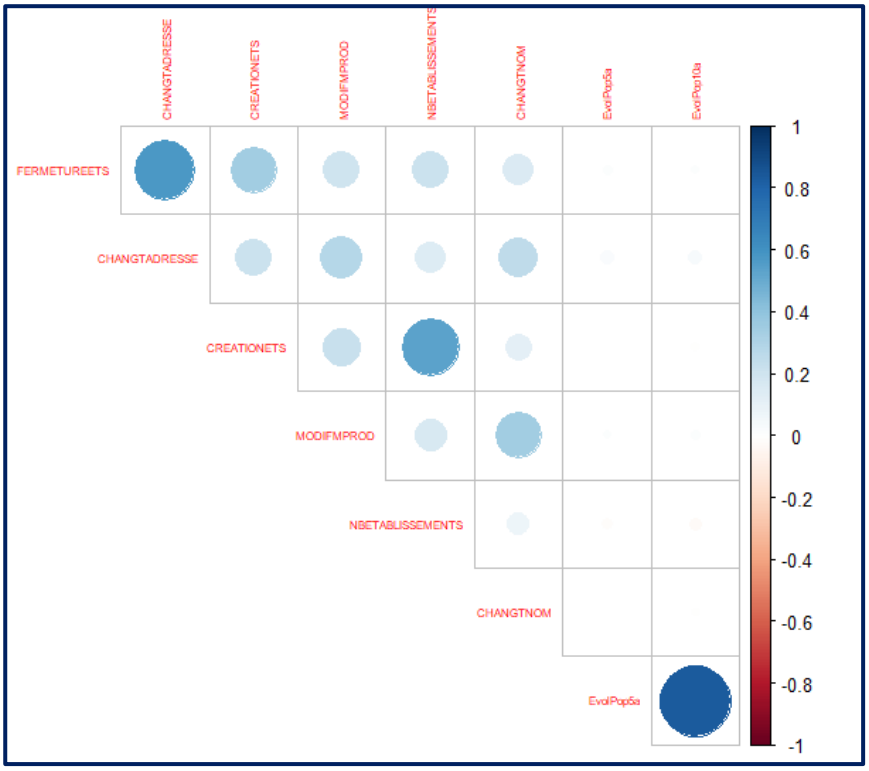
Où $\overline{R_X}$ et $\overline{R_Y}$ sont les moyennes des rangs $R(X_i)$ et $R(Y_i)$.

Le choix sera ici fait d'éliminer les variables dont le coefficient de corrélation est supérieur à 0,75. Le graphique suivant permet de visualiser les variables pour lesquelles une corrélation non négligeable (coefficient supérieur à 0,5) a été décelée.

2.3.2.3.1 Analyse bi-variée des variables financières



Analyse bi-variée des variables non financières



Exemples de variables fortement corrélées :

Variable 1	Variable 2	**Coefficient de Spearman**
EFFECTIF	CA	0,8
EFFECTIF	TOTALBILAN	0,7
LIQUIDITEREDUITE	BFR	0,8
TOTALBILAN	CA	0,9
RESULTATEXPLOITATION	RESULTATNET	0,8
CAPITAUXPROPRES	RESULTATEXPLOITATION	0,8
VA_CA	EBE	1,0
EBE_CA	EBE	1,0
CAPITAUXPROPRES	TOTALBILAN	0,8
BFR	VarBFR	-1,0
DELAIECOULEMENTSTOCKS	BFRJOURSCA	0,7
DELAIECOULEMENTFOURNISSEURS	RESULTATEXPLOITATIONCA	-0,6
CHANGEMENTADRESSE	FERMETUREETS	0,5
EvolPop5a	EvolPop10a	0,8

Certains coefficients de corrélation tombent sous le sens car la première variable est une des composantes de la seconde. Par exemple, il est normal que le montant des capitaux propres soit corrélé avec le total du bilan. De même, le résultat d'exploitation correspond à quelques éléments près au résultat net.

De même, certains autres résultats ne sont pas des surprises car ils ne font que confirmer ce que l'ensemble des connaissances en analyse financière énonce déjà, par exemple le lien étroit entre la rentabilité d'exploitation (Résultat d'exploitation/ Chiffre d'affaires) et le DPO (days payable outstanding ou délai d'écoulement des fournisseurs). En effet, plus la rentabilité d'une entreprise est dégradée, plus elle consommera de cash, notamment en ayant recours au crédit fournisseurs, c'est-à-dire en étirant son délai de règlement.

Globalement, cette étape de l'étude n'aboutit pas à des résultats très surprenants, d'autant plus que certaines variables redondantes auraient très bien pu être « filtrées » dès le début de manière arbitraire et simplement par recours au jugement professionnel (EBE, Valeur ajoutée, Résultat d'exploitation, Résultat net sont par exemple quatre métriques proches).

2.3.3 Description du modèle implémenté sous R avec la bibliothèque H2O

2.3.3.1 Présentation de la bibliothèque H2O

L'objectif général est ici d'élaborer un modèle de prévision par classification des individus (SIREN) dans la catégorie Défaillante (1) ou Non Défaillante (0). La modélisation par régression logistique va aboutir à une valeur correspondant à la probabilité de défaillance estimée du modèle.

Concrètement la modélisation a été réalisée avec le logiciel R, en faisant appel aux algorithmes du package H2O. Le package H2O sous R est communément utilisé pour effectuer les régressions sur des variables binomiales (c'est le cas ici puisque la variable à prédire prend la valeur 1 ou 0).

H2O présente l'avantage dans le contexte présent de fonctionner par pénalisation. En effet, cette approche de la régression logistique est particulièrement adaptée lorsque le nombre de variables est très élevé (ce qui est le cas présent avec une cinquantaine de variables qualitatives et quantitatives à traiter), induisant un risque de surapprentissage trop important (*source : https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/glm.html#regularization*) .

Il a été vu plus haut (dans la partie 2.1)), que la régression logistique repose sur le modèle suivant, où le vecteur $A = \{a_0, a_0, \dots, a_k\}$ décrit les coefficients associés à chacune des variables :

$$\pi\,(x) = P\,(Y = 1 \mid X = x) = \frac{1}{1 + e^{-A.x}}$$

$$= S(A.x)$$

où la fonction S est appelée fonction logit.

Les coefficients sont estimés par la maximisation de la vraisemblance sur un échantillon d'apprentissage (ou d'entraînement)

$$A^* = argmax(LV(A)) = argmax\left[\sum_{i=1}^{n} y_i \ln\big(S(A.x)\big) + (1 - y_i) \ln\big(1 - S(A.x)\big)\right]$$

La pénalisation est une méthode consistant à ajouter à ce problème d'optimisation de la vraisemblance un terme supplémentaire. L'objectif ne sera plus seulement la maximisation de la vraisemblance mais de la rendre la plus élevée possible compte tenu de la contrainte de pénalisation (*F. Planchet , 2017*).

L'équation s'écrit désormais de la manière suivante :

$$A^* = argmax[LV(A) - \lambda r(A)]$$

où $r(w)$ est la fonction de pénalisation ; il s'agit d'une distance à l'origine.

L'algorithme de régression logistique sous H2O utilise la méthode d'*Elastic Net Regularisation* qui combine les deux termes les plus usuels de pénalisation :

- Le Ridge L_2 qui consiste à limiter la valeur de certains coefficients. Le terme de pénalisation est défini dans ce cas comme le carré de la norme 2 du vecteur *A :*

  $$A^* = argmax[LV(A) - \lambda \|A\|_2^2]$$

  Avec $l^2 = \|A\|_2^2 = \sum_{i=0}^{k} a_i^2$

- Le LASSO L_1 (*Least Absolute Shrinkage and Selection Operator*) qui consiste à limiter le nombre de variables en forçant certains coefficients à prendre la valeur 0. Ici, le terme de pénalisation est défini comme la norme l^1 du vecteur *A :*

$$A^* = argmax[LV(A) - \lambda\|A\|_1]$$

Avec $l^1 = \|A\|_1 = \sum_{i=0}^{k}|a_i|$

L'Elastic Net combine donc L_1 et L_2 de la manière suivante :

$$\boldsymbol{A^* = argmax\left[LV(A) - \lambda_1\|A\|_1 - \lambda_2\|A\|_2^2\right]}$$

La valeur des paramètres $\lambda_1$et λ_2 sont déterminés directement par H2O grâce à une méthode de validation croisée entre l'échantillon d'entraînement et l'échantillon de test, le but étant de maximiser l'AUC.

Pour cela deux bases de données vont être créées, l'une pour l'apprentissage, l'autre pour le test, ce qui permettra d'évaluer la qualité de chaque modèle sur des individus différents de la base d'apprentissage. Pour ce faire, le choix d'une clé de séparation de 0,75 est fait : les ¾ des individus seront conservés dans l'échantillon d'apprentissage et le ¼ restant seront réservés au test (*cf extrait de code en Annexe 1)*.

2.3.3.2 Analyse du modèle et interprétation des résultats

La bibliothèque H2O sous R livre un certain nombre de métrique en plus des coefficients du modèle. L'ensemble de ces métriques permet l'évaluation de la qualité du modèle ainsi que la comparaison entre plusieurs modèles. Dans cette étude, les métriques suivantes seront plus particulièrement analysées :

- **La MSE (*Mean Squared Error)***

La MSE est la moyenne des erreurs quadratiques. L'erreur est mesurée comme la distance entre le point (réalité) et la droite de régression. La MSE représente donc à la fois la variance et le biais de la statistique.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

De plus, la MSE donne davantage d'importance aux grandes erreurs. Plus l'erreur est grande, plus la MSE est pénalisée. Une MSE faible signifie que le modèle est performant.

- **Le R^2**

Dans un modèle de régression, le R^2, aussi appelé coefficient de détermination, est une mesure la qualité de l'ajustement par rapport aux données. Cette statistique indique le pourcentage de la variance de la variable dépendante (Y = défaillance dans le modèle) expliqué par les variables explicatives.

$$R^2 = \frac{Variance\ expliquée\ par\ le\ modèle}{Variance\ totale\ de\ la\ variable\ à\ expliquer}\ .$$

En règle générale, il est considéré que plus le R^2 est élevé (proche de 1), plus le modèle correspond aux observations. Cependant, un niveau de R^2 faible n'est pas toujours un problème. En effet, certains champs d'études, notamment tout ce qui est relatif au comportement humain, ont des R^2 souvent très faibles car il est justement très difficile de prévoir les comportements humains. Cela ne veut pas dire que les données ne sont pas statistiquement significatives et des coefficients significatifs peuvent toujours représenter correctement les variations moyennes de la variable à expliquer.

- **L'AUC (*Area Under the Curve*):**

Egale à l'aire sous la courbe ROC. Plus l'AUC est proche de 1, plus le modèle est considéré comme étant de bonne qualité.

Le principe est d'abord de classifier les individus selon la prédiction obtenue grâce au modèle. La courbe ROC (*Receiver Operating Characteristic)* est une représentation graphique de la performance des prédictions pour tous les seuils de classification. Le seuil de classification est la probabilité $\widetilde{\mathbb{P}}$ qui va séparer les prédictions en **DEFAILLANT** ou **NON DEFAILLANT** dans le contexte présent. Par exemple : au-dessus du seuil $\widetilde{\mathbb{P}}$ = 0,5 toutes les entreprises seront classées comme

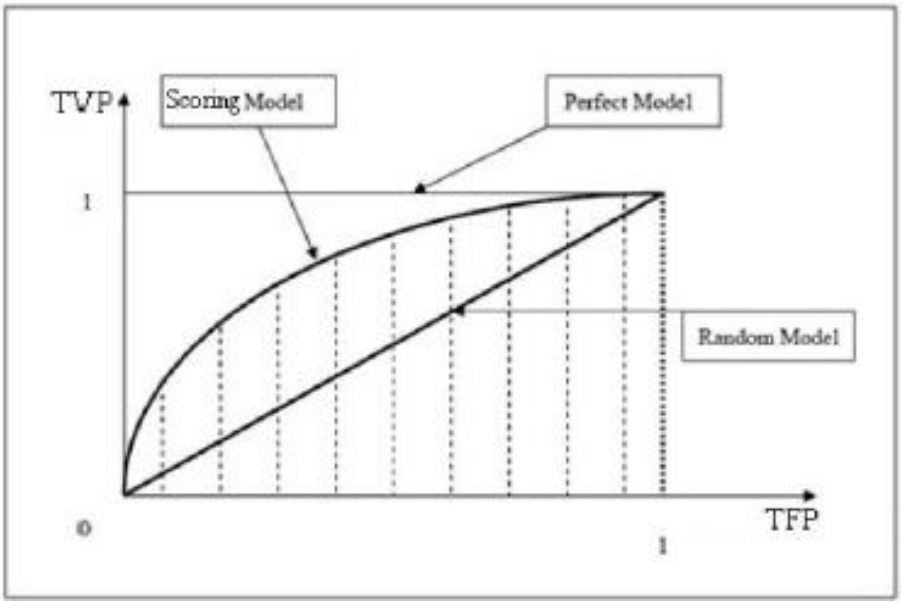
défaillante et en dessous comme non défaillante. Ensuite, une comparaison est faite entre les prédictions et les réalisations de l'échantillon, en déterminant le taux de vrais positifs et le taux de faux positifs :

- Le taux de vrais positifs (TVP) est défini ainsi : $TVP = \frac{VP}{VP+FN}$ où *VP* = vrais positifs et *FN* = faux négatifs
- Le taux de faux positifs (TFP) est donné par : $TFP = \frac{FP}{FP+VN}$ où *FP* = faux positifs et *VN* = vrais négatifs

La **la matrice de confusion** suivante peut d'abord être dressée :

	REALISATION = DEFAILLANTE	REALISATION = NON DEFAILLANTE
PREDICTION = DEFAILLANTE	VP	FP
PREDICTION = NON DEFAILLANTE	FN	VN

Pour tracer le courbe ROC on fait varier le seuil de 1 à 0 et dans chaque cas, après avoir calculé le TVP et le TFP on reporte leur valeur sur le graphique.



Source: Elizabeth MAYS and Niall LYNAS (2010), "Credit scoring for risk managers"

On peut distinguer deux cas extrêmes :

- Dans le premier cas, le modèle n'a aucune qualité prédictive particulière : il est représenté par la droite passant par les points (0,0) et (1,1), soit la courbe y=x. Cela signifie qu'un modèle totalement aléatoire (« random model ») aurait obtenu les mêmes TFP et TVP quel que soit le seuil choisi.
- Dans le second cas, le modèle prédit parfaitement les entreprises qui vont être défaillantes. Quel que soit le taux de faux positifs obtenu, le taux de vrais positifs est 100%, c'est le « perfect model ».

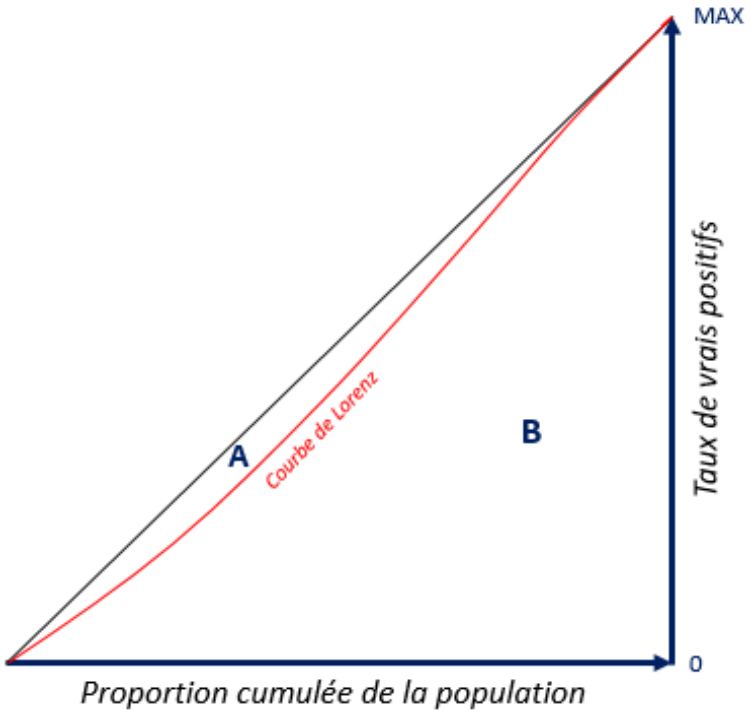
En pratique, la situation sera toujours intermédiaire, c'est-à-dire qu'on obtiendra une courbe qui se situera entre celle du « perfect model» et celle du « random model ». L'exploitation de ce graphique se fait en calculant la mesure AUC (« *area under the curve* »), qui est l'aire entre la courbe et l'axe des abscisses, et dont la valeur, comprise entre 0 et 1 donne la proportion de défaillances correctement détectées par le modèle.

Plus l'AUC est proche de 1, plus le modèle est bon. On considèrera qu'il est satisfaisant si l'AUC est supérieur ou égal à 0,7 et excellent si l'AUC est supérieur à 0,9.

- **Le coefficient Gini** :

Il s'agit d'une statistique particulièrement utile dans un contexte de classification. Il s'agit d'une méthode de quantification de l'hétérogénéité parmi une distribution. Un coefficient de Gini égal à 0 exprime une égalité parfaite (ou encore une classification totalement inutile) et un coefficient de Gini égal à 1 une hétérogénéité maximale (classification parfaite). Ce coefficient se déduit lui aussi de la courbe ROC (comme l'AUC).

Le coefficient de Gini est basé sur la courbe de Lorenz qui décrit le taux de vrai positif en tant que fonction de la proportion de la population classée positive. Il est égal au rapport de la zone A ci-dessous avec la zone A + B.



2.4 Résultats de la modélisation de la défaillance par régression logistique

2.4.1 Coefficients et importance relative des variables explicatives

Les régressions GLM effectuées sur les bases de données brutes, c'est-à-dire sans retraitements sur les variables sélectionnent de manière très radicale les variables significatives, qu'elles soient qualitatives ou quantitatives (discrètes ou continues). En effet, très peu de variables (et donc de modalités de variables pour celles qui sont qualitatives) sont considérées comme significatives :

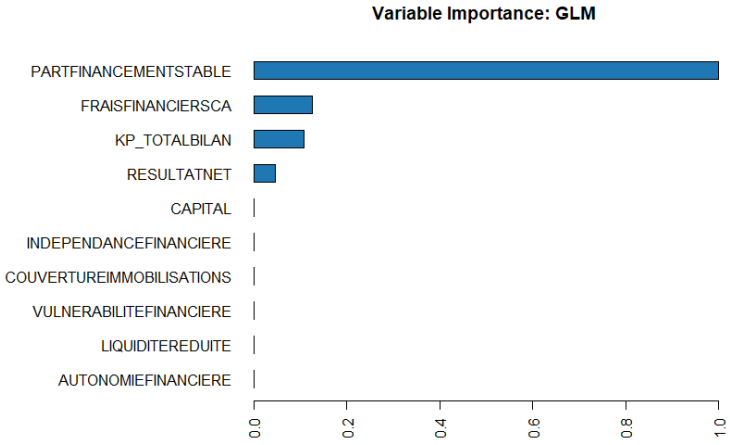
- 4 variables pour le modèle *Data_F* : FRAISFINANCIERSCA, RESULTATNET, KP_TOTALBILAN, PARTFINANCEMENTSTABLE
- 4 variables pour le modèle *Data_ensemble* : DEFAILLANCE, FRAISFINANCIERSCA, RESULTATNET, KP_TOTALBILAN, PARTFINANCEMENTSTABLE

- 22 variables pour le modèle *Data_NF*

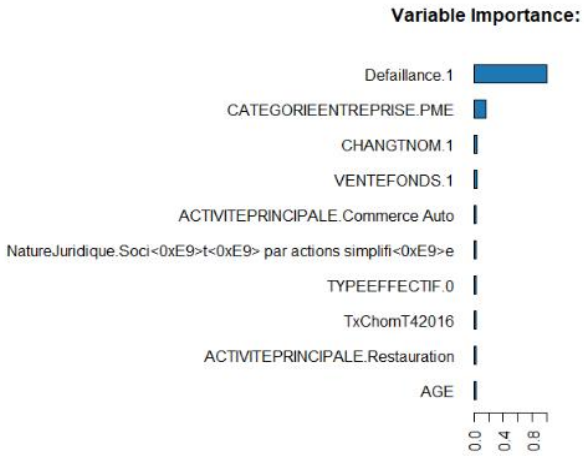
Cf Annexe 2 pour les coefficients détaillés

Au-delà du peu de variables prises en compte dans le modèle, ce qui est le plus frappant est la prépondérance (voire l'hégémonie) d'une variable par rapport aux autres dans chaque modèle. Le graphique suivant illustre l'importance relative de chaque variable dans les résultats du modèle :

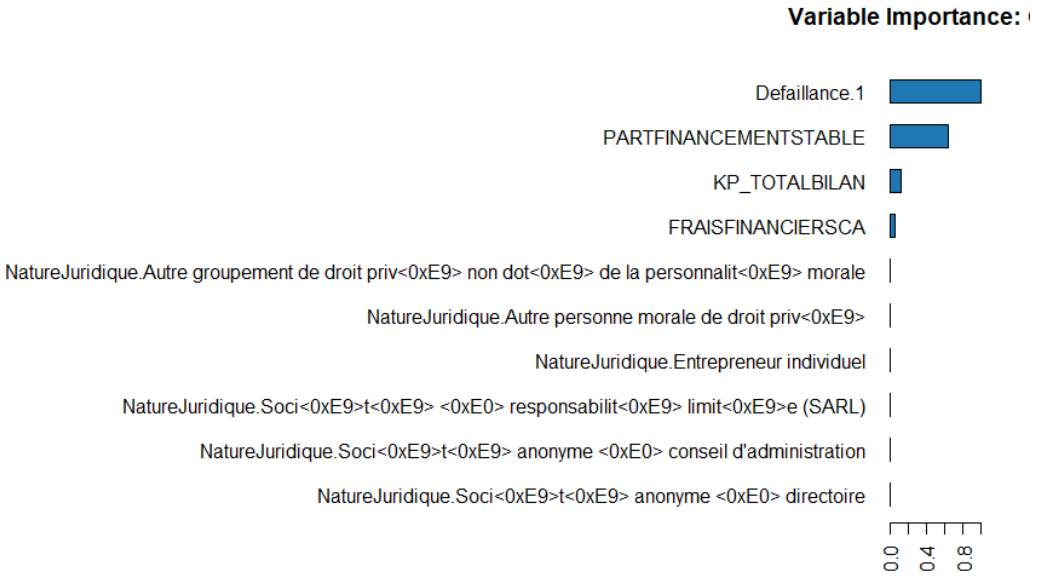
Modèle *Data_F*



Modèle *Data_NF*



Modèle *Data_ensemble*



Ces graphiques illustrent bien l'importance d'une seule variable dans chacun des modèles. Pour les modèles faisant appel aux données non financières (*Data_NF* et *Data_ensemble*), y compris les annonces légales, il s'agit de la variable DEFAILLANCE. Cette variable binaire indique si une annonce de défaillance a été recensée pour le SIREN au cours des 4 dernières années (2013-2016). Dans le cas du modèle *Data_NF,* la seconde variable la plus importante (CATEGORIEENTREPRISE = PME) n'a plus qu'une importance relative de 20 sur une base 100 par rapport à cette variable DEFAILLANCE. L'importance de la variable DEFAILLANCE était attendue car dès l'établissement des statistiques univariées, il apparaissait qu'environ 45% des SIREN ayant eu une annonce de défaillance dans les 4 dernières années, subissait de nouveau une défaillance en 2016 (pour une fréquence empirique générale proche de 1%).

Les coefficients du modèle *Data_F* sont moins attendus car là aussi, c'est une seule variable qui explique la plus grande partie des résultats : la variable PARTFINANCEMENTSTABLE. Dans les statistiques descriptives (partie 2.3.1 font apparaitre la significativité de nombreuses variables (des ratios financiers), ce qui semblait confirmer la pertinence des méthodes traditionnelles d'analyse financière. Le fait que seul le ratio des financements stables soit conservé par le modèle (ou presque…) semble donc indiquer que la quasi-totalité des informations apportées par les autres ratios sont contenues dans ce dernier.

Quant au modèle *Data_ensemble,* il consacre bien l'importance de la variable DEFAILLANCE tout en conservant l'apport du ratio PARTFINANCEMENTSTABLE.

2.4.2 Performances absolues et relatives des différents modèles

Le tableau suivant récapitule les principaux indicateurs de performance des trois modèles ici testés.

	Variables financières uniquement	*Variables non financières: données signalétiques + annonces légales + données exogènes*	*Variables financières et non financières*
	Data_F	**Data_NF**	**Data_ensemble**
MSE	0,013	0,008	0,010
R^2	-0,005	0,192	0,011
Gini	0,509	0,562	0,710
AUC	0,755	0,781	0,855

Les statistiques MSE et R^2 sont à des niveaux peu satisfaisants. En particulier, le modèle *Data_F* affiche des piètres résultats (R^2 négatif en particulier ce qui signifie que des meilleurs résultats auraient été obtenus par le hasard).

En revanche, l'index de Gini et l'AUC qui sont plus adaptés aux objectifs de classification (contrairement au MSE et au R^2 qui sont de bons indicateurs lorsque l'objectif est la prédiction de probabilités) sont bien plus encourageants.

Au global, il peut être affirmé sans trop de doutes que le modèle obtenant les meilleurs résultats est le modèle basé sur l'ensemble des données, aussi bien du point de vue de l'AUC que du point de vue de Gini. Le MSE et le R^2 semblent toutefois donner un léger avantage au modèle basé sur les données non financières uniquement.

2.4.3 Analyse critique et remédiation

2.4.3.1 Les classes déséquilibrées engendrent une instabilité du modèle

Dans cette étude, il y a un déséquilibre flagrant entre les deux classes de la variable à expliquer. En effet, la modalité 1 de la variable DEFAILLANCE est extrêmement rare par rapport à la modalité 0.

(dans la base test)	**Base *Data_F***	**Base *Data_NF***	**Base *Data_ensemble***
nombre d'individus	56 641	73 446	46 498
DEFAILLANCE = 1	676	802	479
DEFAILLANCE = 0	55 965	72 644	46 019
Fréquence DEFAILLANCE	1,19%	1,09%	1,03%
rapport d'occurrence	*83*	*91*	*96*

Ainsi, par exemple dans la base *Data_F* il y a environ 83 SIREN prenant la modalité 0 (non défaillance durant l'année 2017) pour 1 SIREN prenant la modalité 1. Les variables déterminantes de la défaillance ont donc des difficultés à être correctement évaluées. Dans l'absolu, ce n'est pas tant le déséquilibre des classes que le faible nombre d'événements à prédire (nombre de SIREN prenant la valeur DEFAILLANCE = 1 limité à une centaine) qui pose problème ici.

De même, lorsque le programme est de nouveau lancé sous R, la séparation de la base de données originale en deux (75% des individus pour la base d'entraînement et 25% des individus pour la base de test) donne des répartitions sensiblement différentes. La présence d'un SIREN défaillant de plus dans la base d'entraînement va induire des résultats potentiellement variables.

2.4.3.2 Des données peuvent être biaisées ou trop complexes

Le fait que les fréquences empiriques de défaillance ne soient pas identiques après les différents retraitements effectués, alors même que les travaux portent sur un échantillon identique (mais en tenant compte de variables différentes) doit amener à s'interroger sur les biais introduits par l'échantillon lui-même. En effet, après retraitements, la fréquence

empirique des défaillances est de 1,19% pour la base de données *Data_F* et de 1,09% pour la base *Data_NF.* Il semble donc probable que l'information relative à la présence ou non des données (publications des comptes, exhaustivité des données signalétiques…) soit en elle-même une indication importante pour la prédiction de la défaillance, et que sa non prise en compte induirait un biais significatif.

2.5 Amélioration des modèles initiaux

2.5.1 Modélisation par régression avec catégorisation des données au préalable

2.5.1.1 Présentation des retraitements effectués

Les résultats de la régression logistique semblent en contradiction avec l'analyse univariée réalisée dans la partie 2.3.1. En effet, une grande partie des variables semblaient à première vue être significatives dans le cadre de l'analyse de la fréquence des défaillances. Dans ce cadre, un certain nombre de variables quantitatives présentes dans la base de données des entreprises ayant publié (les variables financières) avaient été rendues discrètes en les divisant en quantiles (en l'occurrence des déciles).

Des régressions logistiques ont été réalisées sur l'ensemble des données prises de manière brutes ce qui permettait d'avoir dans le même modèle des variables catégorielles et des variables quantitatives.

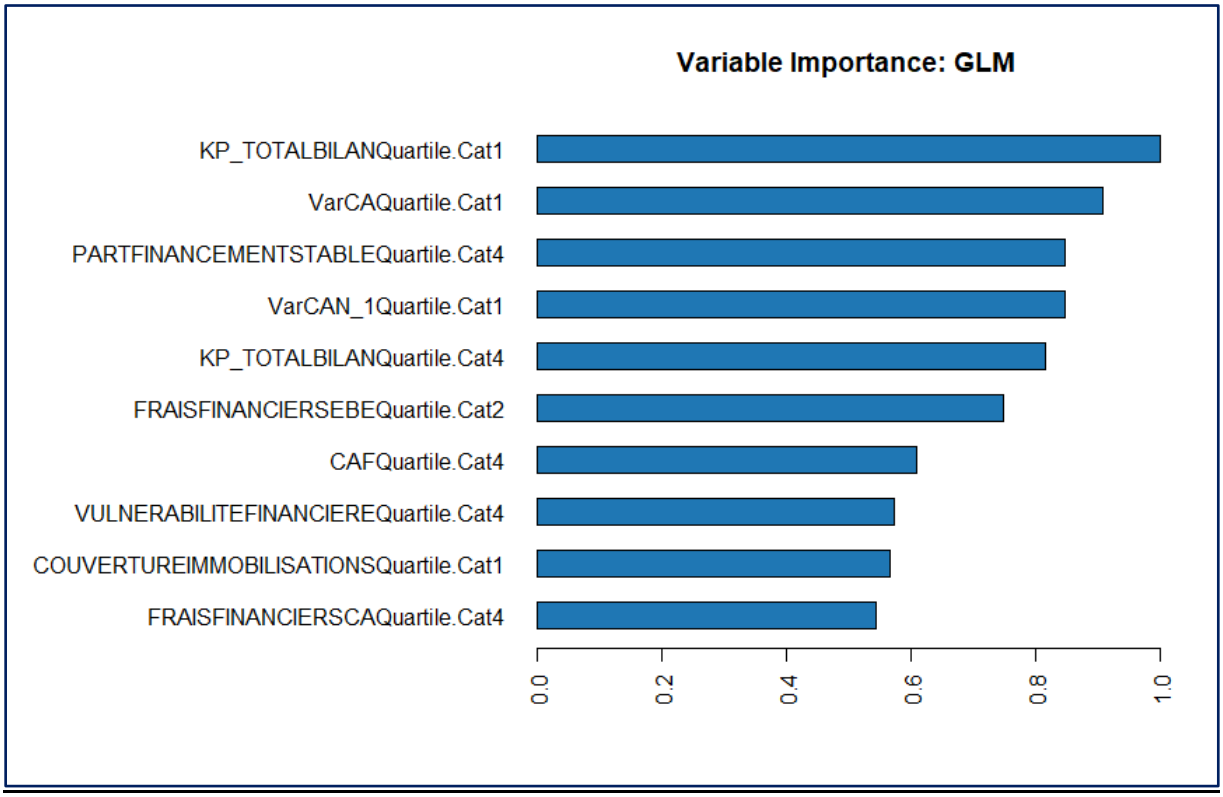
Dans cette partie, les variables quantitatives (discrètes et continues) de la base de données ont été modifié pour les rendre qualitatives avant de refaire tourner le modèle. Ces modifications ont concerné principalement les données financières (qui sont continues). La catégorisation de ces variables continues a consisté à les séparer en quartiles, chaque quartile devenant une modalité de la variable.

D'autres retraitements sont également intervenus sur les variables signalétiques suivantes : regroupement de l'ancienneté (AGE) en trois modalités seulement, et concentration des codes NAF en quatre secteurs : hôtellerie (codes naf commençant par 55), restauration (codes naf commençant par 56), commerce de détail automobile (codes naf commençant par 47), commerce de détail hors automobile (codes naf commençant par 45).

2.5.1.2 Résultats des régressions logistiques sur des données catégorisées

2.5.1.2.1 Une meilleure prise en compte des variables introduites dans le modèle

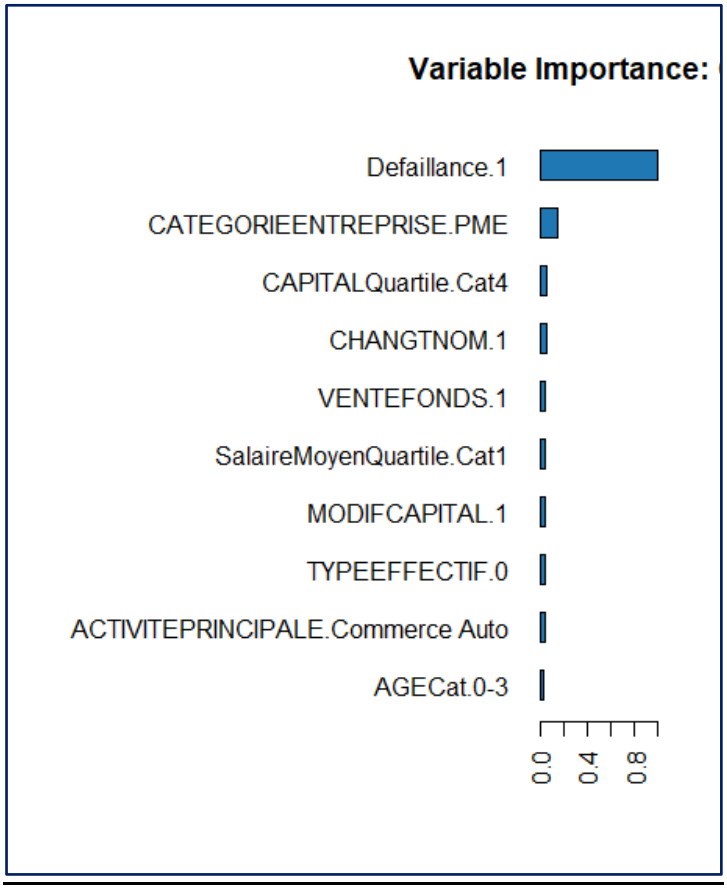
- **<u>Graphique de l'importance relative des variables : modèle *Data_F*</u>**



Ce graphique illustre l'importance relative des 10 variables prépondérantes, ce qui permet de réaliser à quel point le retraitement des données effectué au préalable, équilibre l'importance des variables dans le modèle GLM sur *Data_F.* Du point de vue de l'expert, ces résultats ont beaucoup plus de sens, car l'ensemble de ces ratios sont pris en compte dans une analyse financière traditionnelle. En outre, plusieurs ratios semblent avoir leur importance et apporter leur part d'information, contrairement au modèle sur base non retraitée (non catégorisée).

En détail, ce sont 65 variables qui ont un impact non nul dans ce modèle, dont 34 à impact positif et 31 à impact négatif (coefficients associés respectivement positif et négatif). *Cf Annexe 2.*

- **<u>Graphique de l'importance relative des variables : modèle *Data_NF*</u>**



Les quelques retraitements effectués sur les données signalétiques ne semblent pas avoir eu d'impact substantiel sur l'importance relative des variables. La présence d'une annonce légale de défaillance antérieure l'emporte toujours largement par rapport aux autres variables. En revanche, le nombre de variables significatives a augmenté (il est passé de 21 à 38).

2.5.1.2.2 Performance des modèles retraités

	Variables financières uniquement		*Variables non financières: données signalétiques + annonces légales + données exogènes*		*Variables financières et non financières*	
	Data_F		**Data_NF**		**Data_ensemble**	
	après retraitements	*avant retraitements*	*après retraitements*	*avant retraitements*	*après retraitements*	*avant retraitements*
MSE	0,013	0,013	0,008	0,008	0,008	0,010
R^2	0,041	-0,005	0,201	0,192	0,233	0,011
Gini	0,732	0,509	0,576	0,562	0,851	0,710
AUC	0,866	0,755	0,788	0,781	0,926	0,855

L'analyse des scores de ces modèles est sans appel : les 4 indicateurs pris en compte (MSE, R^2, Gini, AUC) sont en amélioration non négligeable. Seule la MSE des modèles *Data_F* et *Data_NF* ne s'améliore pas (mais ne se dégrade pas non plus).

Les résultats pour le modèle sur données financières uniquement sont en amélioration très significative : l'AUC passe de 0,755 à 0,866, le Gini de 0,509 à 0,732 et le R^2 de -0,005 à 0,041.

Contrairement aux premiers résultats obtenus avant les retraitements effectués sur les variables, désormais de meilleurs résultats sont obtenus en se basant sur les données financières uniquement qu'en ne tenant compte que des données signalétiques, les annonces légales et les données exogènes. En effet, les 4 indicateurs étaient meilleurs pour *Data_NF* que pour *Data_F* avant retraitement. A l'exception du R^2 qui reste meilleur pour le modèle *Data_NF,* les autres indicateurs sont désormais à des meilleurs niveaux pour *Data_F.*

En outre, c'est le modèle *Data_ensemble*, utilisant l'ensemble des données disponibles (financières et non financières) qui dépasse les deux autres en termes de résultats. Cela signifie que toutes les catégories de variables apportent leur part d'information et se complètent.

Ainsi, bien que meilleur théoriquement, le modèle *Data_F* ne sera pas utilisé, dans aucun des cas de figure. En effet, il convient de rappeler que deux types de situations existent en réalité dans l'étude des défaillances d'entreprises :

- Une première dans laquelle les éléments financiers ne sont pas disponibles (du fait d'une qualité de données insuffisante ou de la non-publication des états financiers)
- Une seconde dans l'ensemble des éléments financiers et extra financiers sont disponibles

Dans la première situation, ni le modèle *Data_F,* ni le modèle *Data_ensemble* ne pourront être appliqués. Seul le modèle *Data_NF* pourra être appliqué.

Dans la seconde situation, le modèle utilisé sera *Data_ensemble* car il reste meilleur que *Data_F.*

2.5.2 Retour à une base non biaisée dans le cas d'une indisponibilité des variables financières

2.5.2.1 La base de données des entreprises ayant publié leurs compte présente dès le départ un léger biais

Dès le début de l'étude une base de données unique a été utilisé communément pour toutes les situations. En effet, dans l'optique de tester l'efficacité d'un modèle sans données financières par rapport à un modèle basé sur des données financières (approche « classique »), l'ensemble des données disponibles dans les systèmes d'information du groupe Pouey ont été extrait, pour l'ensemble des Siren des secteurs analysés (commerce de détail, hôtellerie, restauration) ayant publié leurs comptes en 2014, 2015, 2016.

L'efficacité de différentes catégories de données a ainsi pu être testée : données financières versus autres (signalétiques, annonces et exogènes) dans le cadre d'un modèle GLM, pour les mêmes individus.

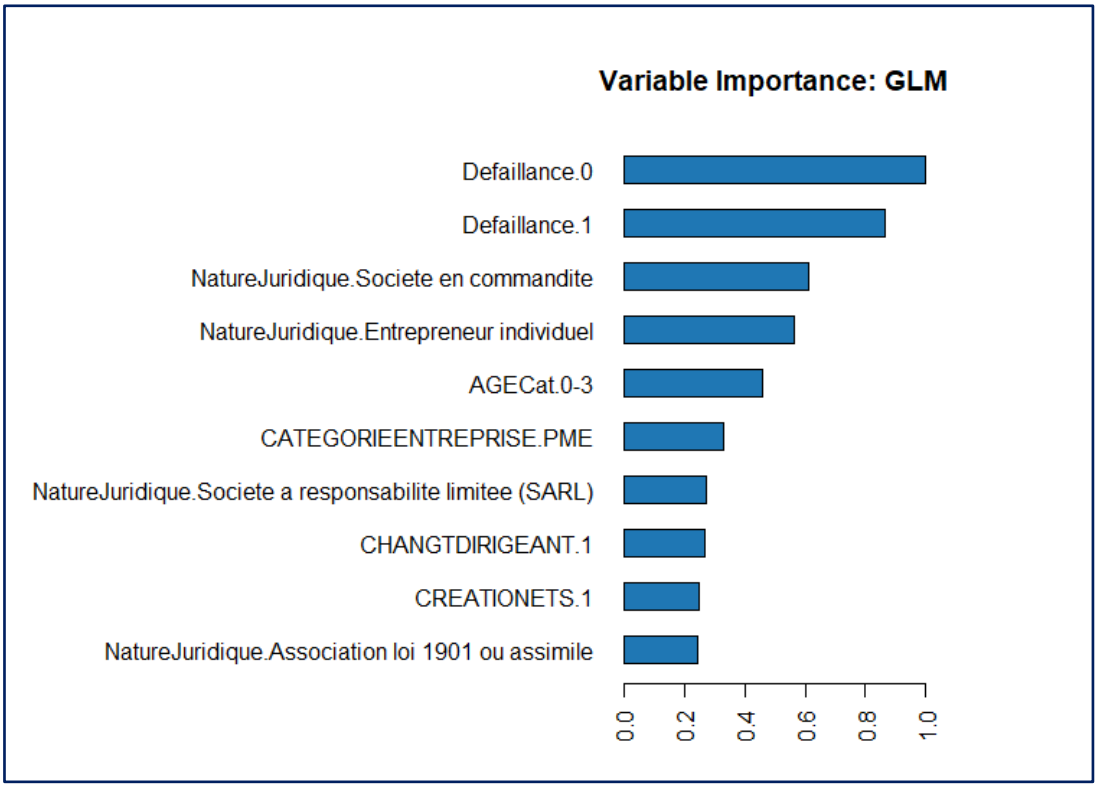
Cela ne tient pas compte d'un biais qui pourrait sembler anodin au premier abord : le fait de publier ou de ne pas publier ses comptes n'est-il pas une indication en soit ? Ainsi, en ne travaillant sur une base de données ne contenant que des entreprises ayant publié, n'y a-t-il

pas un risque d'être soumis à ce biais qui rendrait en fait le modèle applicable aux entreprises n'ayant pas publié leurs comptes caducs ?

Pour répondre à la première interrogation, il a été nécessaire de revenir à la base de données la plus large possible avec pour seul critères l'appartenance aux secteurs concernés (codes naf commençant par 45,47, 55, 56) et le fait que l'entreprise (SIREN) soit encore active. Après « nettoyage de la base de données » contenant plus d'un million de SIREN actifs, environ 340 000 lignes ont pu être conservée, ce qui est un progrès significatif par rapport à la base de données initiale qui en contenait environ 18 000. Quant à la fréquence empirique des défaillances dans cet échantillon, elle est de 1,85% contre 1,03% dans la base des entreprises ayant publié leurs comptes entre 2014 et 2016. Cela signifie donc bien que le fait de publier ses comptes est une information en soit et que l'entraînement du modèle uniquement sur des entreprises ayant publié leur compte introduirait un biais.

2.5.2.2 Des résultats en amélioration et d'une plus grande stabilité

- **Graphique de l'importance relative des variables dans le modèle *Data_NF*, sur la base exhaustive des entreprises actives**



Ce graphique illustre une prise en compte plus équilibrée des variables dans le modèle. La variable de défaillance antérieure reste la variable la plus importante mais ne contient plus l'intégralité de l'information du modèle comme c'était le cas (ou presque) dans la base des entreprises ayant publié leurs états financiers. Cela peut s'expliquer par le nombre plus important d'entreprises ayant fait défaillance dans l'échantillon (plusieurs milliers), ce qui permet au modèle de dégager des tendances liées à d'autres variables. Il parait également probable que ce modèle sera plus robuste et stable grâce à sa taille et le nombre d'occurrence de l'événement prédit.

Quant aux statistiques du modèle, elles sont les suivantes :

	Data_NF	
	BDD comptes 2014-2016 disponibles	***BDD exhaustivité des entreprises actives***
MSE	0,008	0,017
R^2	0,201	0,091
Gini	0,576	0,707
AUC	0,788	0,854

Globalement le modèle est plus fiable à l'aune de ces statistiques : AUC et Gini sont en amélioration par rapport au modèle basé sur l'échantillon des entreprises ayant publié leurs comptes entre 2014 et 2016. En revanche, il convient de noter que la MSE et le R^2 sont à des moins bons niveaux mais ces indicateurs sont à privilégier surtout dans une optique de précision de la prédiction et moins dans le cadre d'une classification.

Le choix sera donc fait d'utiliser les résultats de ce modèle qui est plus adapté à un échantillon non biaisé.

3. Une modélisation alternative : utilisation de la méthode Random Forest

3.1 Présentation de la méthode Random Forest

L'algorithme Random Forest (ou Forêts Aléatoires) est une méthode proposée en 2001 par *L. Breiman* consistant à agréger plusieurs arbres de décisions décorrélés et en ne conservant plus que la moyenne de leurs résultats. Cette méthode est particulièrement recommandée lorsque le nombre de variables est important et les événements à prédire rares.

Les principaux atouts des arbres de régression sont :

- la simplicité de leur compréhension et de leur interprétation
- leur performance dans un environnement de grands jeux de données (nombre d'individus élevé et nombre de variables explicatives élevées)
- ils fonctionnent avec des variables à la fois quantitatives et qualitatives
- par rapport à la régression linéaire, la relation entre les différents X_i et Y ne doit pas nécessairement être linéaire
- il n'y a pas d'hypothèse de normalité obligatoire

L'objectif de la méthode est in fine d'estimer une fonction de régression du type

$$\boldsymbol{m(x) = E[Y|X = x]} \text{ avec}$$

- $X = \left(X^{(1)}, \dots, X^{(p)}\right)$ le vecteur des variables explicatives
- $Y = \left(Y_1, \dots, Y_p\right)$ la variable à expliquer

3.1.1 Prérequis : les arbres de décision CART

Le principe de l'arbre de décision de type CART (*Classification And Regression Tree*) est de pouvoir classer les individus dans une catégorie, en créant des « feuilles » les plus homogènes possibles par rapport à la variable à prédire (*C. Chesneau, 2019)*. Ces feuilles sont

issues de ramifications, qui elles même découlent d'un processus de séparation des individus en fonctions des variables et des seuils permettant cette séparation-là plus homogène possible.

La détermination de la valeur (ou modalité s'agissant d'une variable qualitative) la plus plausible est celle qui est la plus observée parmi les individus d'une feuille.

A chaque étape, la variable et le seuil qui divisent l'échantillon de manière à minimiser l'erreur de prédiction sont choisis.

En termes mathématiques, l'objectif est de déterminer j_* et c_* qui rendent minimal la valeur $G\ (j, c)$, appelée indice de Gini défini de la manière suivante :

Les variables explicatives de Y sont $X_1, X_2, \dots, X_p$

Soit $j \in \{1, \dots, p\}$, c un nombre réel. A chaque nœud on a :

- $n_{j,c,gauche}$ le nombre de SIREN vérifiant $X_j < c$
- $n_{j,c,droite}$ le nombre de SIREN vérifiant $X_j > c$
- $f_{1,j,c,gauche}$ est la fréquence des défaillances pour les SIREN vérifiant $X_j < c$ (et de manière identique on définit $f_{0,j,c,gauche}$ pour les SIREN non défaillants)
- $f_{1,j,c,droite}$ est la fréquence des défaillances pour les SIREN vérifiant $X_j > c$ (et de manière identique on définit $f_{0,j,c,droite}$ pour les SIREN non défaillants)
- $G_{gauche}(j, c) = 1 - ({f_{1,j,c,gauche}}^2 + {f_{0,j,c,gauche}}^2)$
- $G_{droite}(j, c) = 1 - ({f_{1,j,c,droite}}^2 + {f_{0,j,c,droite}}^2)$

L'ensemble de ces valeurs sont ici définies pour des variables quantitatives mais il en est de même pour les variables qualitatives en remplaçant la séparation par le seuil *c* (et donc la création de deux intervalles) par des sous-ensembles des différentes modalités de chaque variable explicative).

On mesure ainsi la perte d'information globale lors de la séparation des SIREN en deux groupes de la manière suivante :

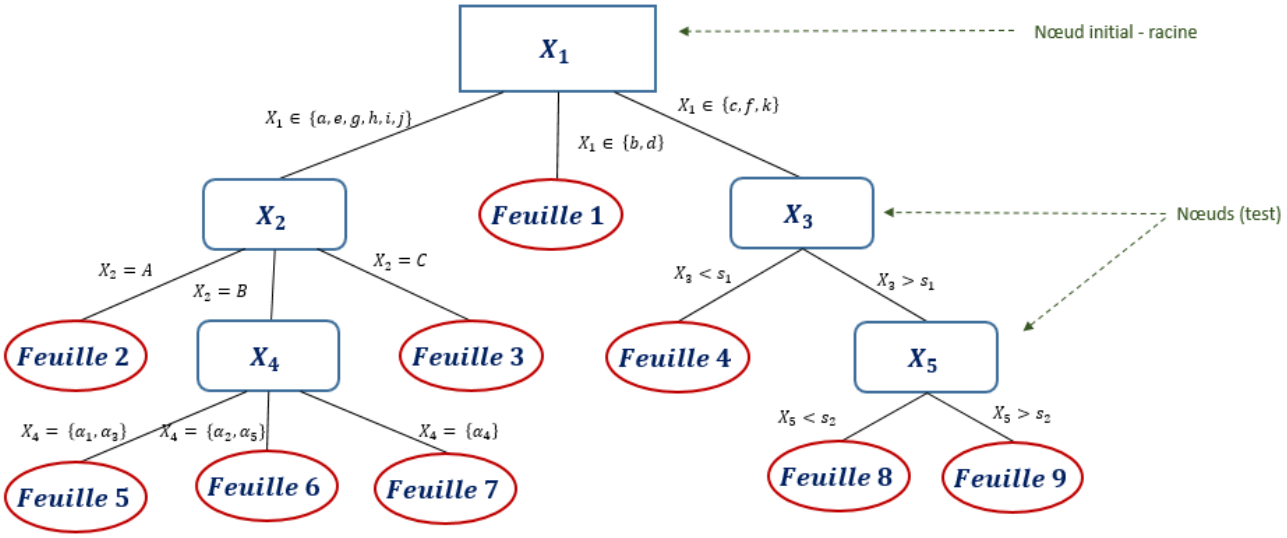
$$\boldsymbol{G\ (j, c) = \frac{n_{j,c,gauche}}{n} . G_{gauche}(j, c) + \frac{n_{j,c,droite}}{n} . G_{droite}(j, c)}$$

Lors de l'établissement d'un arbre de décision, on peut fixer des conditions qui limiteront l'étendue de la propagation de l'arbre :

- Une profondeur maximale : nombre de ramification maximales que peut prendre une branche
- Un nombre d'individu minimal par feuille.

Si l'un de ces deux critères n'est pas fixé, le processus peut potentiellement aller jusqu'à ce qu'il ne reste plus qu'un seul individu par feuille, ce qui induirait évidemment un fort risque de suraprentissage.

L'observation de la modalité majoritaire dans chaque feuille permet l'utilisation de l'arbre à des fins de classification.



3.1.2 Les forêts aléatoires d'arbres de décision

Le but de la méthode des forêts aléatoires est de palier à l'instabilité des arbres de décision. Dans cette méthode une construction aléatoire d'arbres est réalisée pour ensuite les combiner ensemble et utiliser le résultat issu du calcul de la moyenne des arbres. Pour réaliser cela, les 3 paramètres suivants sont choisis :

- Nombre d'individus sélectionnés aléatoirement

- Nombre de variables sélectionnées aléatoirement
- Nombre d'arbres générés
- Critère de limitation des arbres de décisions (voir ci-dessus)

Une fois la forêt d'arbres aléatoire construite, il suffit de placer les individus pour lesquels on souhaite réaliser une prédiction sur chacun de ces arbres et d'en faire la moyenne.

3.1.3 Evaluation de l'importance des variables dans le cadre de Random Forest

Contrairement aux modèle GLM, la classification Random Forest ne délivre pas de coefficients pour chaque variable. Ainsi il faut trouver une autre méthode pour évaluer l'importance relative de chaque variable.

Pour ce faire, à chaque fois qu'une variable est choisie pour créer une nouvelle branche dans un arbre de décision, l'erreur quadratique moyenne est calculée avant et après, ce qui permet d'attribuer une amélioration du modèle lié à cette variable. Une fois la forêt aléatoire réalisée, la somme des améliorations pour chaque variable est calculée, puis mise sur une échelle commune (ci-dessous entre 0 et 1).

3.2 Classification du risque de défaillance par la méthode Random Forest

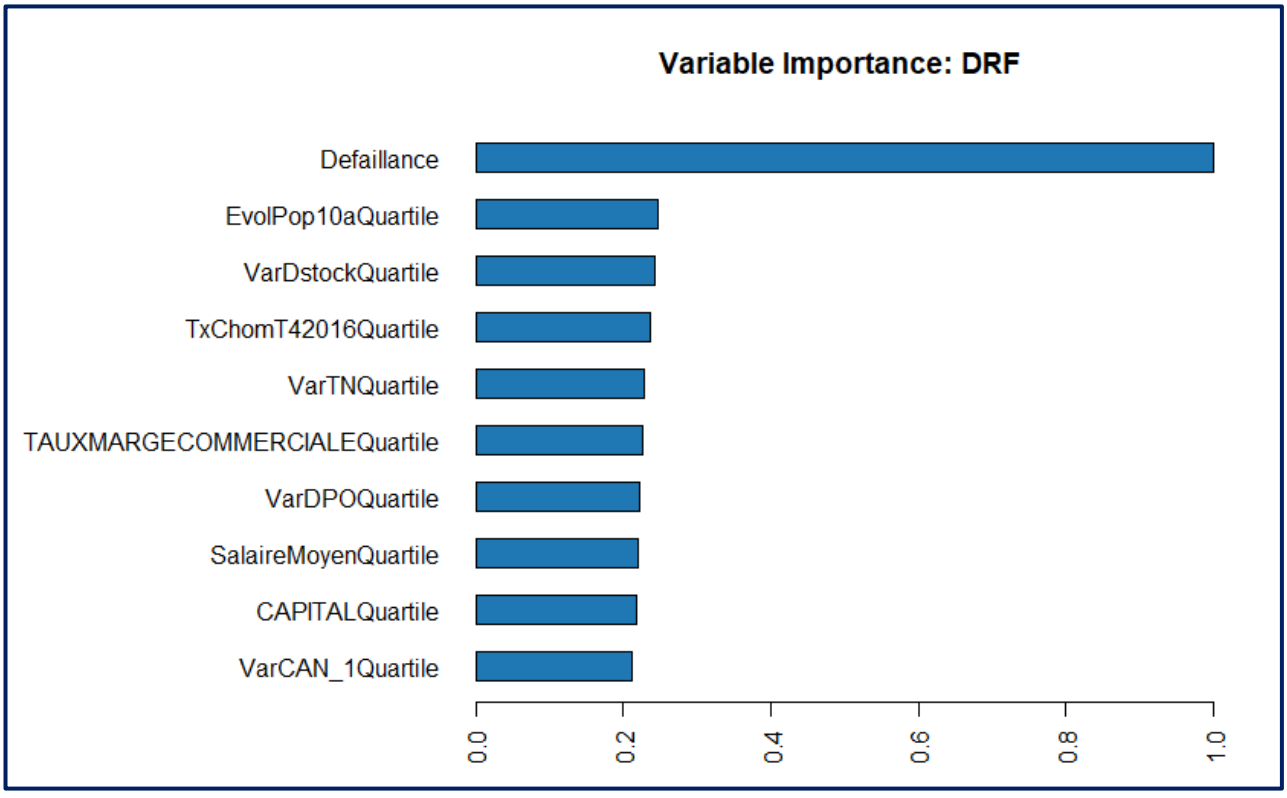
	BDD exhaustivité des entreprises actives		*BDD comptes 2014-2016 disponibles*	
	Data_NF		**Data_ensemble**	
	GLM	***Random Forest***	***GLM***	***Random Forest***
MSE	0,017	0,016	0,008	0,004
R^2	0,091	0,116	0,233	0,600
Gini	0,707	0,766	0,851	0,935
AUC	0,854	0,883	0,926	0,967

La qualité des modèles de forêts aléatoires par rapport aux modèles GLM est indéniable : sur les mêmes bases d'entraînement et de test, toutes les statistiques sont meilleures avec la méthode des forêts aléatoires.

Il est probable que l'interactivité entre les variables, prise en compte dans les forêts aléatoires à l'inverse de la GLM, est déterminante pour la prédiction des défaillances.

L'amélioration est particulièrement spectaculaire pour le modèle *Data_ensemble* : Gini passe de 0,851 à 0,967 et l'AUC plafonne à 0,967 (contre 0,926).

- **<u>Graphique de l'importance relative des variables dans le modèles Random Forest sur *Data_ensemble*</u>**



4. Validation des modèles et interprétation dans une optique de tarification

4.1 Application des meilleurs modèles régression logistique et Random Forest dans un contexte d'assurance-crédit

4.1.1 Echantillon d'entreprises ayant fait l'objet d'une cotation en 2016 par Pouey International

Dans les parties précédentes, les modèles ont été entraînés et leur pertinence testée sur les bases de données les plus larges et exhaustives possible. En premier lieu le travail a porté sur les bases de données de toutes les entreprises des secteurs Hôtellerie, Restauration et Commerce de détail pour lesquelles les données financières (issues des publications de comptes) étaient disponibles et exploitables. Mais il est vite apparu que le fait de disposer des données financières était en soi une première indication.

Pour vérifier cela, les travaux se sont focalisés sur la base exhaustive des entreprises actives au 31/12/2016 : la fréquence empirique des défaillance l'année suivante (2017) sur cette base était de **1,85% contre 1,05%** dans la base des entreprises ayant publié leurs comptes entre 2014 et 2016.

Ces deux bases exhaustives et non biaisées (selon la situation avec ou sans données financières disponibles) ont permis l'entraînement de modèles valables pour des cas aléatoires. Qu'en est-il dans la réalité, c'est-à-dire si l'objectif est d'appliquer ces modèles aux portefeuilles d'entreprises ayant fait l'objet d'une cotation par Pouey International ? L'application de ces modèles à ces portefeuilles va-t-elle permettre d'établir un tarif par l'utilisation de la matrice de confusion ?

4.1.2 Limite imposée par la taille réduite des portefeuilles de simulation

Ces bases « portefeuilles » sont restreintes en termes de quantité de SIREN. Ainsi, en 2016, ce sont 1963 cotations (dites « enquêtes ») qui ont été réalisées par Pouey International dans ces secteurs d'activité. Mais seulement 346 enquêtes ont porté sur des SIREN pour lesquels les données financières sont disponibles

Le nombre d'occurrence de l'événement à prédire (la défaillance en 2017) est beaucoup trop restreint : 130 occurrences dans l'ensemble des enquêtes, mais seulement 7 pour les SIREN avec des données financières disponibles.

En outre, ces enquêtes ne sont pas toujours réalisées dans une optique de garantie. En isolant les enquêtes pré-garanties (dites « Preseren » et « Sélection »), le nombre d'occurrence de la défaillance tombe à 61 pour les SIREN présentant les données non financières uniquement, et à 3 sur les SIREN pour lesquels les données financières sont disponibles.

- **Récapitulatif du nombre d'individus par échantillon d'enquêtes**

défaillances 2017/enquêtes 2016	**Toutes enquêtes**		**Prégaranties**	
	données non financières uniquement	***avec données financières***	***données non financières uniquement***	***avec données financières***
Nombre d'individus	1617	346	686	131
Nombre de défaillances	123	7	61	3
Fréquence empirique	**7,6%**	**2,0%**	**8,9%**	**2,3%**

4.1.3 Les portefeuilles de Pouey International sont biaisés

La fréquence empirique des défaillances dans ces échantillons est largement supérieure à celle constatée dans les bases d'entraînement. En effet, la base des enquêtes avec disponibilité des données financières fait apparaître une fréquence empirique d'environ 2,30% contre 1,05% dans la base d'entraînement. Pour les SIREN sans données financières, l'écart est encore plus important : la fréquence passe de **1,85%** sur la base d'entraînement **à 8,89% sur la base des**

enquêtes pré-garanties. Il apparait donc que le portefeuille de cotations pré-garanties de Pouey International subit manifestement une antisélection très importante.

Il convient ici de donner une explication à cette antisélection. Les cotations (enquêtes) de Pouey International sont réalisées en partie dans une optique de garantie (assurance-crédit). L'enquête réalisée par Pouey International (et la cote / 20 qui en ressort) détermine si une garantie peut être accordée sur cette entreprise et à quel tarif (en fonction du niveau de la cote). Le positionnement de marché de Pouey International pour commercialiser ces garanties implique qu'une grande partie de ces enquêtes ont déjà fait l'objet d'un refus par une autre entité d'assurance-crédit. En outre, les nombreux outils d'analyse et de surveillance des portefeuilles clients, disponibles sur le marché, permettent aux entreprises de faire leur propre sélection des clients risqués.

Ainsi, il parait impossible de fonder un tarif d'assurance-crédit sur les résultats du modèle appliqués au portefeuille d'enquêtes réalisées par Pouey International.

4.2 Test du modèle sur une base re-échantillonnée

4.2.1 Application des modèles et comparaison avec les résultats de test

Pour remédier au faible nombre d'entreprises présentes dans le portefeuille des enquêtes réalisées, ainsi qu'au biais introduit par l'antisélection de cette base, les modèles retenus vont être appliqués sur des nouvelles bases de données, les plus larges possible, en les re-échantillonnant pour reproduire les conditions de l'antisélection décrites plus haut.

4.2.1.1 Pour les SIREN sans données financières : bases des entreprises actives au 31/12/2017

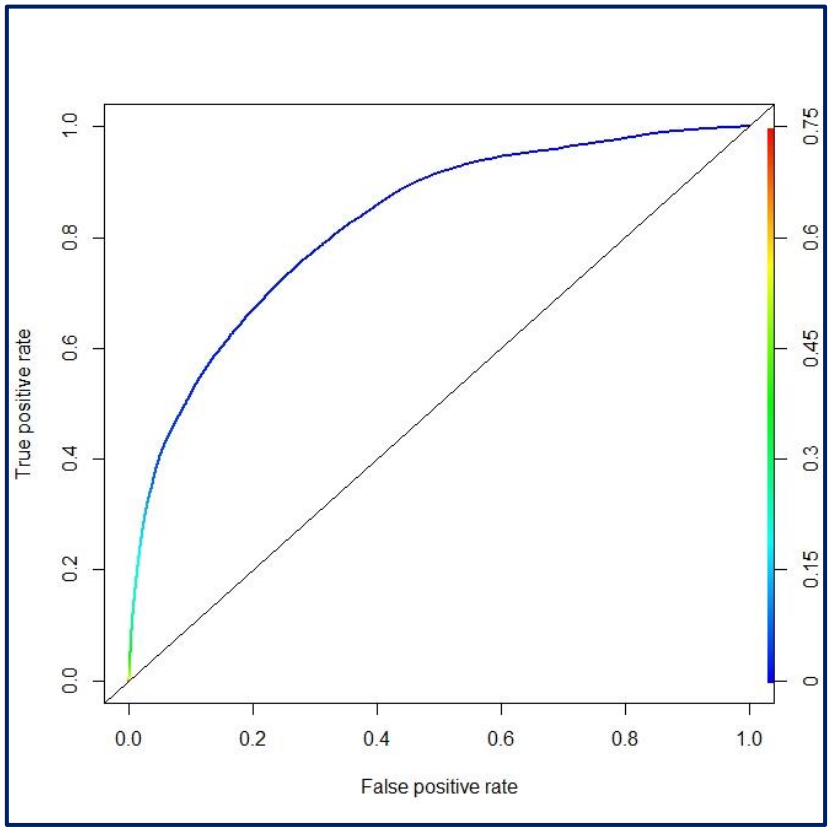
Pour le modèle *Data_NF*, les travaux ont d'abord porté sur la base de données des entreprises des secteurs étudiés actives au 31/12/2017. Cette base de données présente le double avantage de ne pas avoir été utilisée au préalable comme base d'entraînement, et surtout de présenter un nombre d'individu suffisant.

Pour recréer les conditions de l'antisélection et atteindre artificiellement une fréquence empirique des défaillances correspondant à celle des enquêtes pré-garanties pour les SIREN

sans données financières (environ 8,9% pour rappel), la majeure partie des SIREN défaillants (environ 80%) a été conservée et autant de SIREN non défaillants que nécessaires ont été tirés aléatoirement pour arriver à la proportion attendue.

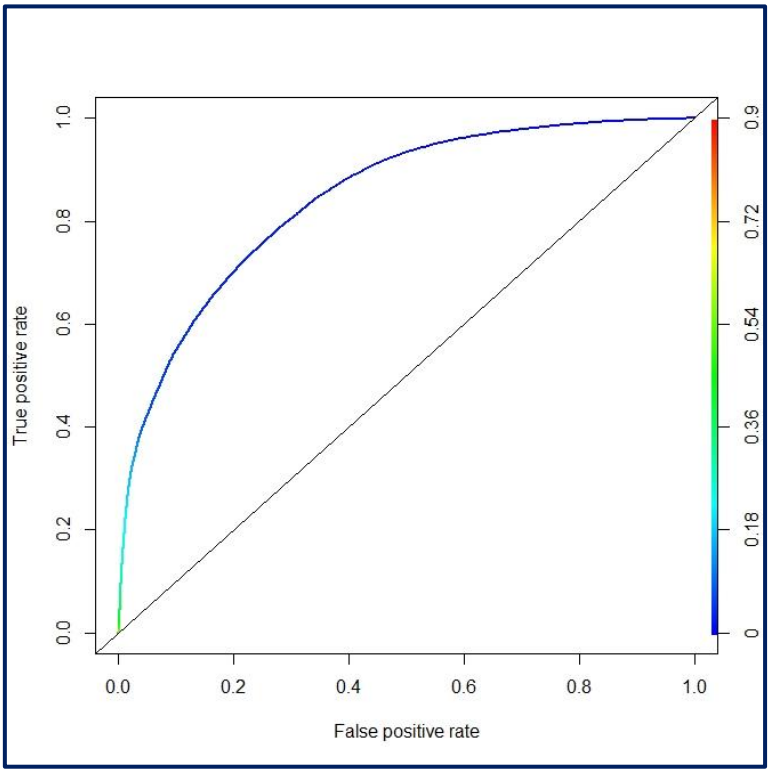
4.2.1.1.1 Résultats obtenus en régression logistique

	Data_NF	
	base non rééchantillonée	*base rééchantillonnée*
MSE	0,017	0,078
R^2	0,091	0,071
Gini	0,707	0,650
AUC	0,854	0,825



4.2.1.1.2 Résultats obtenus avec la méthode Random Forest

	Data_NF	
	base non rééchantillonée	*base rééchantillonnée*
MSE	0,017	0,075
R^2	0,091	0,103
Gini	0,707	0,687
AUC	0,854	0,843

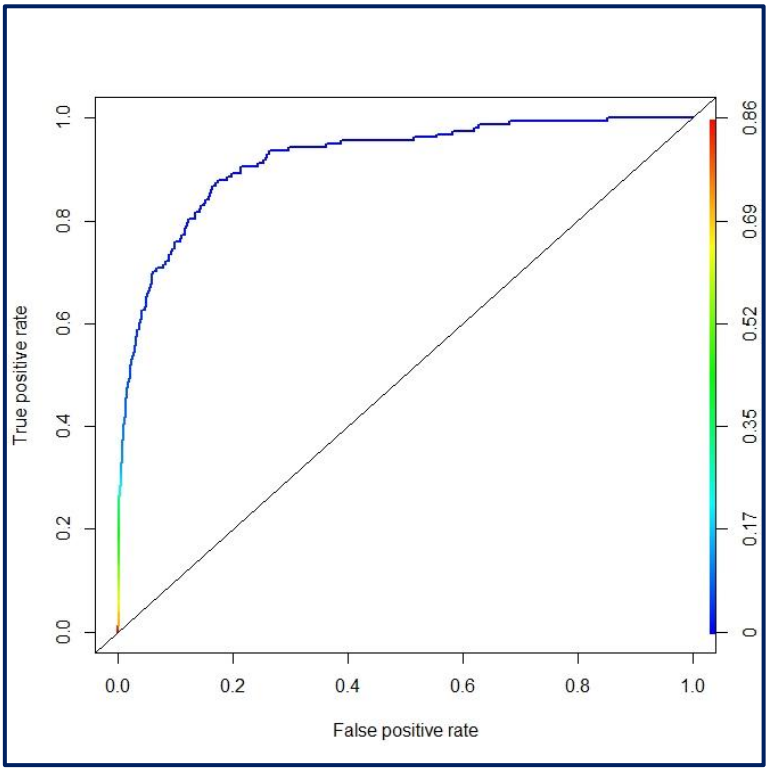


4.2.1.2 Pour les SIREN avec données financières : retour à la base de test initiale (publications 2014-2016)

Concernant les entreprises pour lesquelles les données financières sont disponibles, un re-échantillonnage de la base de donnée *Data_ensemble* qui avait été utilisée comme base d'apprentissage et de test du modèle a été effectué afin d'atteindre une fréquence des défaillances de 2,63%.

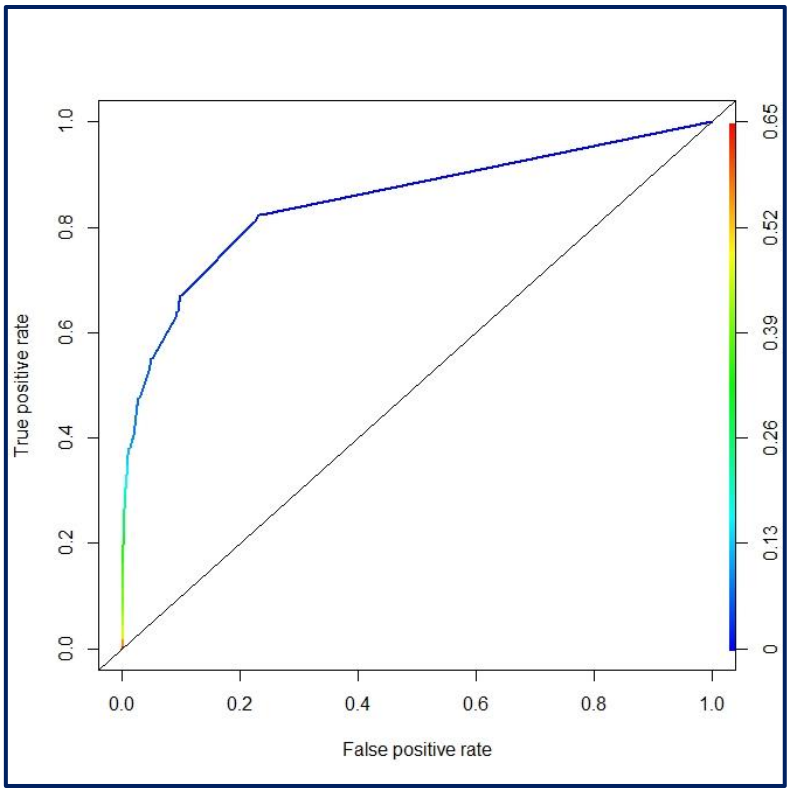
4.2.1.2.1 Résultats obtenus en régression logistique

*comparatif test sur base rééchantillonnée - **GLM***	**Data_ensemble**	
	base non rééchantillonée	***base rééchantillonnée***
MSE	0,008	0,020
R^2	0,233	0,208
Gini	0,851	0,840
AUC	0,926	0,920



4.2.1.2.2 Résultats obtenus avec la méthode Random Forest

*comparatif test sur base rééchantillonnée - **RF***	**Data_ensemble**	
	base non rééchantillonée	***base rééchantillonnée***
MSE	0,004	0,021
R^2	0,600	0,166
Gini	0,935	0,699
AUC	0,967	0,849



4.2.2 Exploitation des résultats par l'analyse de la matrice de confusion

Dans le contexte de cette étude la classification consiste à fixer un seuil entre 0 et 1, au-delà duquel l'entreprise sera considérée comme défaillante et en dessous duquel elle sera prédite comme non défaillante. L'observation des valeurs réelles permettra de tracer la matrice de confusion prédiction VS réalité.

	REALISATION = DEFAILLANTE	**REALISATION = NON DEFAILLANTE**
PREDICTION = DEFAILLANTE	***VP***	***FP***
PREDICTION = NON DEFAILLANTE	***FN***	***VN***

A partir de cette matrice de confusion, les scores suivants peuvent être définis :

- $\boldsymbol{Precision} = \frac{VP}{VP+FP}$
- $\boldsymbol{Recall} = \frac{VP}{VP+FN}$

- $\boldsymbol{Accuracy} = \frac{VP+VN}{Nombre\ d'observations}$
- $\boldsymbol{F1} = 2\left(\frac{(precision)(recall)}{precision+recall}\right)$
- $\boldsymbol{F0,5} = 1{,}25\left(\frac{(precision)(recall)}{0{,}25precision+recall}\right)$
- $\boldsymbol{F2} = 5\left(\frac{(precision)(recall)}{4precision+recall}\right)$
- $\boldsymbol{Mean\ per\ class\ Accuracy} = \frac{1}{2}\left(\frac{VN}{VN+FP} + \frac{VP}{FN+VP}\right)$
- $\boldsymbol{Min\ per\ class\ Accuracy} = Min\left[\frac{VN}{VN+FP}; \frac{VP}{FN+VP}\right]$

Tous ces scores ont leur pertinence en fonction du contexte dans lequel s'inscrit l'étude. Le logiciel R permet, pour chaque score, de trouver le seuil qui le maximise. Ainsi, pour chaque score, un seuil de classification différent sera choisi.

Par exemple, dans le cas du modèle Random Forest appliqué aux données non financières uniquement, les différents seuils calculés sous R et les matrices de confusion correspondantes sont détaillés dans l'*Annexe 3.*

Ces résultats permettent également de comparer les différents modèles entre eux. Par exemple, en prenant comme métrique la plus pertinente la *Min per class Accuracy,* les données sont les suivantes :

			PREVISION = 0	PREVISION = 1	Per class accuracy	Seuil
Data_NF	GLM	**REALISATION = 0**	67 651	23 173	0,745	**0,015**
		REALISATION = 1	2 493	6 751	0,730	
	Random Forest	**REALISATION = 0**	68 557	22 172	0,756	**0,025**
		REALISATION = 1	2 208	6 970	0,759	
Data_ensemble	GLM	**REALISATION = 0**	2 455	518	0,826	**0,010**
		REALISATION = 1	7	66	0,904	
	Random Forest	**REALISATION = 0**	2 609	287	0,901	**0,020**
		REALISATION = 1	30	53	0,639	

4.2.3 Détermination du seuil de classification dans une optique d'assurance-crédit

4.2.3.1 Principe général de l'exploitation des résultats en vue de la tarification

En assurance-crédit, il est possible pour l'assureur de sélectionner les risques acceptés et ceux refusés. L'objectif est de pouvoir baser le tarif sur une fréquence inférieure aux fréquence empiriques observées : ici la fréquence empirique des enquêtes pré-garantie. Cependant, il ne s'agit pas d'être excessivement sélectif : en effet l'intérêt de l'assureur comme de l'assuré est de pouvoir accepter le maximum de risques :

- Pour l'assuré afin d'être couvert sur une part la plus grande possible de son portefeuille de client
- Pour l'assureur afin de pouvoir collecter un maximum de primes

L'enjeu est ici de trouver le seuil qui optimise à la fois :

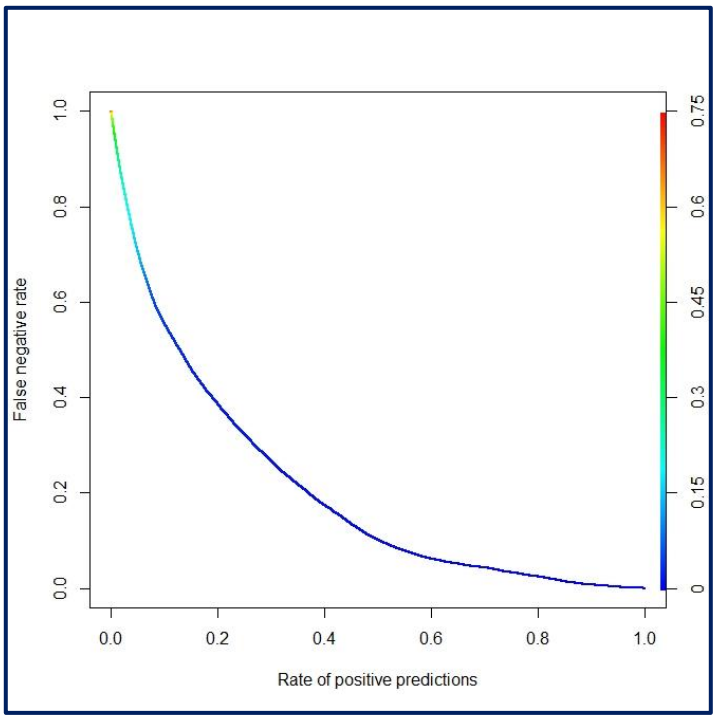
- Le taux de faux négatifs : $TFN = \frac{FN}{FN+VN}$ qui représente le taux de défaillance parmi les SIREN classés non défaillants. Ce taux est incontournable car c'est celui qui sera utilisé dans la détermination du tarif.
- Le taux de faux positifs : TFP = $\frac{FP}{FP+VP}$. Il s'agit du taux d'entreprises classifiées comme défaillantes à tort. Ces entreprises vont faire l'objet d'un refus, alors qu'elles auraient pu être couvertes sans risque et générer une collecte de prime plus importante.

L'événement DEFAILLANCE = 1 étant beaucoup plus rare que l'inverse, pour diminuer le TFN il faut abaisser le seuil de classification de manière relativement importante, ce qui nécessairement va faire basculer une part non négligeable de SIREN non défaillants dans la catégorie DEFAILLANTS et ainsi augmenter le taux de refus.

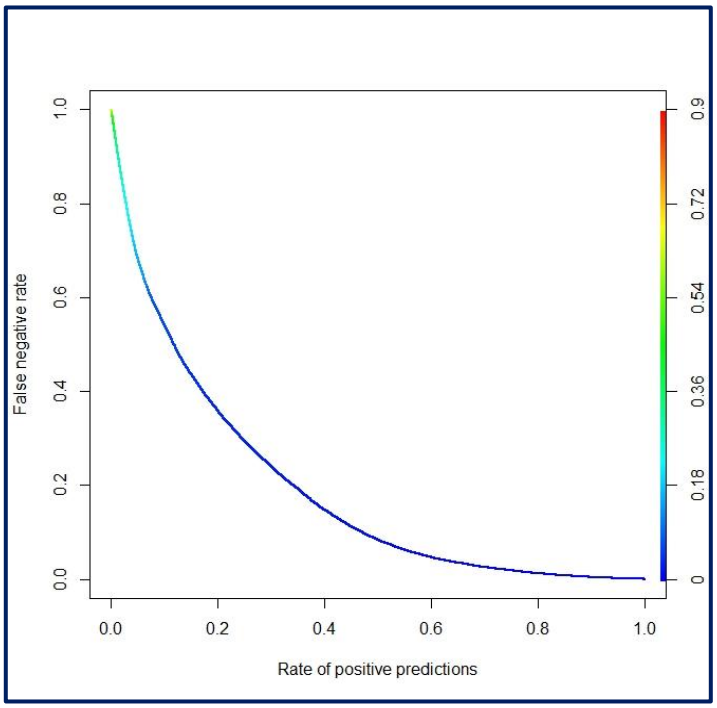
Dans cette optique, il est donc pertinent d'observer le graphique représentant le taux de faux négatifs (« *false negative rate* ») en fonction du taux de faux positifs (« *rate of positive predictions* »). Il va être possible de comparer l'efficacité des modèles les uns par rapport aux autres en fixant un niveau donné pour l'un des deux indicateurs, et en regardant lequel des modèles minimise le deuxième indicateur. Par exemple, pour un TFN de 3% à quel niveau s'établit le TFP pour chacun des modèles. Celui qui aura la résultante la moins élevée pourra être considéré comme le plus performant.

4.2.3.2 Classification des entreprises sans données financières

4.2.3.2.1 Par Régression Logistique

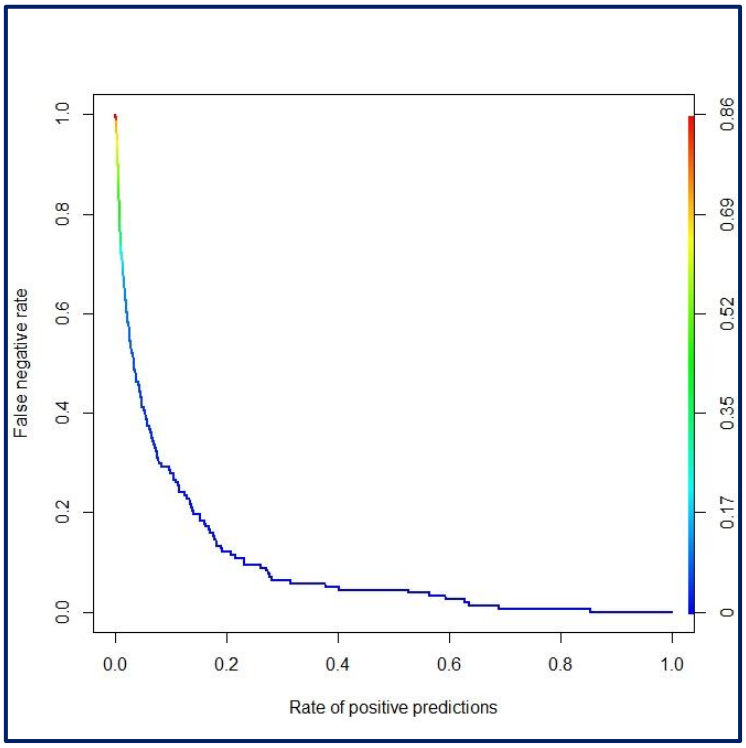


4.2.3.2.2 Par Random Forest

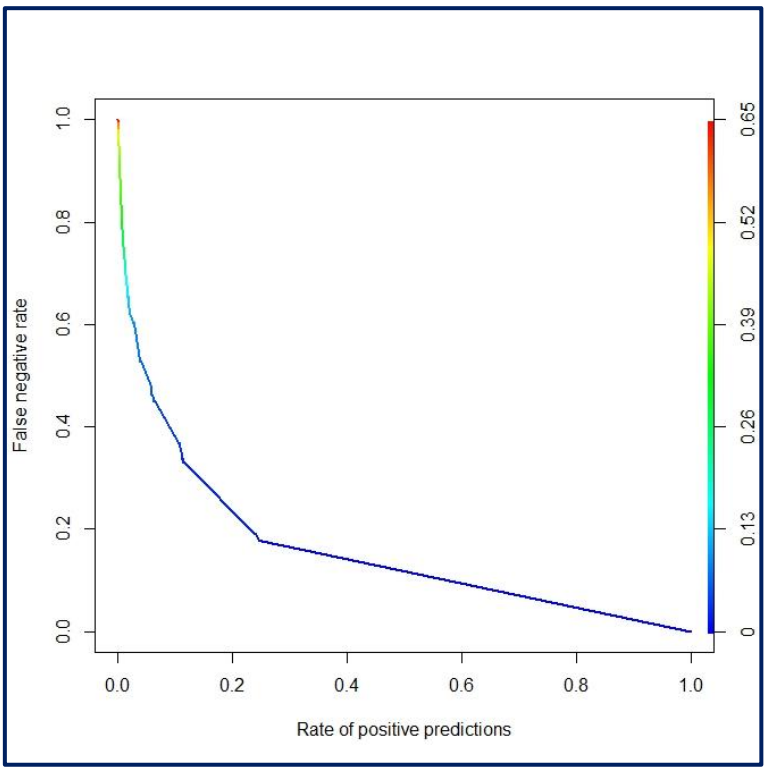


4.2.3.3 Classification des entreprises avec données financières

4.2.3.3.1 Par Régression Logistique



4.2.3.3.2 Par Random Forest



4.2.4 Comparaison de l'efficacité des modèles dans une optique d'assurance-crédit

Dans une optique d'assurance-crédit, il apparait donc clairement que le modèle permettant de minimiser le taux de refus (= $\frac{VP+FP}{VP+FP+VN+FN}$) pour un taux de sinistralité ***TFN*** donné, ou inversement qui minimise le ***TFN*** pour un taux d'acceptation cible, sera le meilleur. Dans l'étude présente, le seuil de classification importe peu car le tarif ne sera pas basé sur la probabilité estimée mais sur le ***TFN***.

- **<u>Cas des modèles basés sur les données non financières :</u>**

Modèle	Taux de refus	Seuil	TFN
Régression logistique	**20,00%**	0,0209	4,44%
Random Forest		0,0342	**4,13%**
Régression logistique	35,96%	0,0127	**3,00%**
Random Forest	**31,68%**	0,0225	

L'avantage de Random Forest sur la Régression Logistique apparait très clairement. En effet, pour un taux de refus fixé à 20,00%, le ***TFN*** avec Random Forest est de 4,13% alors qu'il est de 4,44% avec la Régression Logistique.

Inversement, pour atteindre un ***TFN*** de 3,00%, il faut refuser 35,96% des cas en Régression Logistique, donc plus qu'avec Random Forest (31,68% des cas).

- **<u>Cas des modèles basés sur les données financières :</u>**

Modèle	Taux de refus	Seuil	TFN
Régression logistique	**15,00%**	0,014	**0,61%**
Random Forest		[0,019 - 0,020]	[0,657% - 0,980%]
Régression logistique	**7,83%**	0,028	**0,85%**
Random Forest	[11,53% - 23,93%]	[0,019 - 0,020]	

Il apparaît qu'en présence des données financières, la classification par Régression Logistique est plus efficace que la méthode Random Forest. Pour un taux de refus fixé à 15,00%, le ***TFN***

déduit est de 0,61% en Régression logistique et se situe entre 0,657% et 0,98% en Random Forest.

Inversement, pour atteindre un ***TFN*** de 0,85%, il faut refuser 7,83% des cas en Régression logistique et entre 11,53% et 23,93% en Random Forest.

4.3 Comparaison de l'efficacité des modèles de classification obtenus par rapport aux analyses de Pouey International

La méthode actuelle de cotation des entreprises par analyse « à dire d'expert » revient à réaliser une classification de type binomiale. En effet, les notes attribuées aux entreprises après analyse s'établissent entre 0 et 20. Une entreprise obtenant une note supérieure ou égale à 9 pourra faire l'objet d'une garantie (pour une durée de 12 mois), et inversement celles dont la note est inférieure ou égale à 8 seront refusées.

D'après les données historiques des enquêtes réalisées par Pouey International sur la période 2016 et 2017 dans les secteurs hôtellerie, restauration et commerce de détail, rapprochées des annonces légales de défaillance de l'année suivante (respectivement 2017 et 2018), la matrice de confusion suivante peut être dressée :

	PREVISION = 0	PREVISION = 1
REALISATION = 0	2 735	771
REALISATION = 1	105	156

Ainsi, le modèle d'analyse de Pouey International obtient les métriques suivantes :

TFN	3,70%
TFP	83,17%
Taux de refus	24,61%
Precision	16,83%
Recall	59,77%
Accuracy	76,75%
F1	0,263
F0,5	1,497
F2	0,396
Mean per class Accuracy	68,89%
Min per class Accuracy	59,77%

La méthode de classification de Pouey International permet donc de passer d'une fréquence empirique de 6,93% à un taux de sinistralité (**TFN)** de 3,70%, mais c'est au prix d'une sélection drastique : près d'un quart (24,61%) des demandes refusées, alors que parmi elles, seules 17% seront en défaillance (ce qui veut dire que 83% des demandes refusées auraient pu être garanties).

Il est donc possible de comparer ces résultats obtenus par Pouey International à ceux qui auraient été obtenus en appliquant les modèles étudiés dans ce mémoire. Le meilleur critère de comparaison est simplement de vérifier si le taux de faux négatifs obtenu pour le même taux d'acceptation est plus ou moins élevé :

Modèle	Taux de refus	Seuil	TFN
Régression logistique	~24,6%	0,0233	4,08%
Random Forest		0,0319	2,68%

Ainsi, il apparait très clairement que par Random Forest, pour le même taux de refus, un résultat meilleur est obtenu (TFN égal à 2,68% contre 3,70%) qu'avec les cotations de Pouey International. En revanche, la Régression logistique donne un résultat plus mauvais (TFN égal à 4,08% contre 3,70%).

4.4 Construction d'un produit d'assurance-crédit à partir des modèles retenus

4.4.1 Principes généraux de détermination des niveaux de garanties et du tarif

Les modèles calculés et testés dans ce mémoire ont prouvé une mauvaise qualité de prédiction (MSE élevée et R^2 faibles) mais une capacité de classification beaucoup plus intéressante. Ainsi, il apparait impossible de se baser sur la probabilité individuelle de défaillance sur chaque SIREN pour déterminer le tarif.

En revanche, le ***TFN*** de chaque modèle testé sur des échantillons à la fois volumineux (nombre d'individus importants) et représentatif du portefeuille d'assurance de PRCG (fréquence

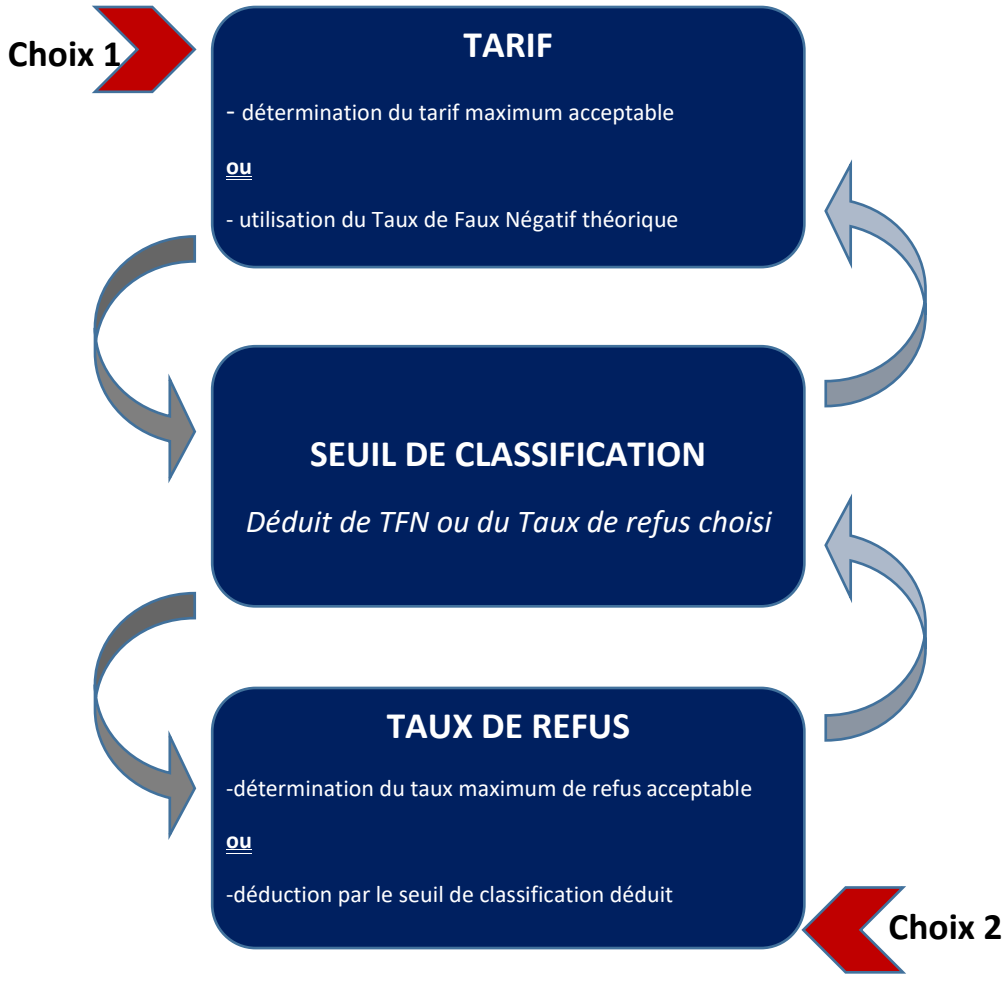
empirique équivalente) est la statistique clé de ce mémoire : c'est sur elle que pourront être fondés les tarifs d'assurance-crédit issus de l'étude (prime pure).

Il va donc être possible de bâtir un produit d'assurance reposant sur ces modèles. Le principe serait de laisser à l'assuré la main sur un des deux leviers suivants :

- la proportion de son portefeuille qu'il souhaiterait garantir (nombre d'entreprise garanties / nombre d'entreprises total du portefeuille)
- le tarif maximum acceptable

Le choix par le client du nombre (ou de la proportion) de refus de garanties toléré de sa part induirait la fixation du seuil de classification adéquat et par construction (projection sur la courbe TFN/Taux de refus) au tarif, correspondant au **TFN**.

De même, la fixation d'un taux de tarif accepté par le client (TFN implicite) induirait la fixation du seuil de classification et donc à un certain nombre de refus de garanties.



4.4.2 Simulation et étude de la sensibilité sur le portefeuille des cotations Pouey International

Comme décrit plus haut, les résultats obtenus vont être utilisés en appliquant les modèles Random Forest et Régression Logistique sur la base de données exhaustive des SIREN, par l'utilisation de la matrice de confusion et du seuil de classification implicite.

Le portefeuille sur lequel la simulation du tarif d'un produit d'assurance-crédit va porter, est le portefeuille des enquêtes réalisées par Pouey International en 2016 et 2017 sur les secteurs d'activité Commerce de détail, Hôtellerie, Restauration. Ce portefeuille contient un nombre N = 3767 de SIREN. Pour obtenir un taux d'acceptation de 90% (i.e un taux de refus de 10%), il convient donc de refuser les 377 SIREN pour lesquels la prédiction est la plus élevée.

- **Régression Logistique :**

Dans le modèle régression logistique, cela revient à fixe un seuil de classification à 0,0395. A ce seuil de classification correspond un taux de faux négatifs (TFN_{tarif}) théorique de 5,61%. Or, en appliquant ce seuil de classification au portefeuille de simulation, le taux de faux négatif ($TFN_{réel}$), soit la sinistralité obtenue a posteriori est de 4,93%.

Il y aurait donc une marge égale à $TFN_{tarif} - TFN_{réel}$ = 0,68%

- **Random Forest :**

Dans le modèle Random Forest, pour classifier 377 SIREN comme défaillants, il faut fixer le seuil à 0,0582, induisant un TFN_{tarif} théorique de 5,44%. La fixation de ce seuil dans le portefeuille aboutirait à un taux de sinistralité $TFN_{réel}$ a posteriori de 3,19%. La marge serait ici extrêmement élevée : $TFN_{tarif} - TFN_{réel}$ = 2,25%.

Le modèle obtenu s'avère donc dans le cas présent du portefeuille des enquêtes de Pouey International très pessimiste et donc aboutie à une tarification surévaluée.

Afin d'expliquer la plus grande efficacité des modèles obtenus grâce à cette méthode sur le portefeuille des enquêtes par rapport à la base exhaustive des SIREN, il est probable supposer

que l'on retrouve plus fréquemment les modalités (pour les variables qualitatives) fortement pénalisées ou que les variables associées à des forts coefficients prennent plus souvent des valeurs élevées (variables quantitatives). D'une part, cette hypothèse est difficilement vérifiable mais elle semble plausible dans la mesure où l'on avait pénalisé artificiellement la base exhaustive des SIREN pour recréer les conditions de l'antisélection. D'autre part, il convient de rappeler que les cotations de Pouey International font souvent suite à des refus de garanties de la part d'un assureur de premier rang (Coface, Atradius, Euler Hermes…) qui a lui-même appliqué ses propres modèles auparavant. Il est donc logique que certaines variables fortement pénalisées par le modèle de ce premier assureur apparaissent plus fréquemment dans le portefeuille de cotations de Pouey International.

Conclusion

Dans le cadre de la conception d'un produit d'assurance-crédit dédié aux secteurs de l'hôtellerie, de la restauration et du commerce de détail, la tarification repose principalement sur l'estimation de la probabilité de défaillance des entreprises couvertes.

La disponibilité de bases de données contenant à la fois un grand nombre de variables explicatives et un volume important d'observations a permis la mise en application des deux méthodes d'apprentissage de données suivantes : régression logistique et forêts aléatoire qui ont rapidement montré une plus grande capacité de classification que de prédiction. En outre, afin de recréer le biais d'antisélection des portefeuilles de garanties de Pouey International, ces modèles ont été entraînés sur des bases de données artificiellement biaisées.

A partir de ces différents modèles de classification du risque de défaillance, l'évaluation de la prime pure d'assurance repose sur l'exploitation de la matrice de confusion et notamment l'arbitrage à réaliser entre le taux de faux négatifs et le taux de refus. En effet, c'est ce taux de faux négatifs qui va constituer la valeur de la probabilité de défaillance retenue pour l'établissement du tarif. Un des grands atouts de cette méthode est la possibilité pour l'assureur ou l'assuré de choisir soit un tarif maximum acceptable, soit un taux de refus (ou un nombre de garanties refusées). En effet, une fois le modèle (régression logistique ou forêts aléatoires) appliqué sur le portefeuille à garantir, la table de correspondance générale seuil de classification → TFN induira le résultat recherché (taux de refus ou taux de tarif).

D'autre part le modèle optimisant au mieux le couple TFN/ Taux de refus sera réputé comme étant le plus efficace. Ainsi, en comparaison des analyses (enquêtes) effectuées par Pouey International il apparait que la méthode des forêts aléatoires ne conduit pas à une perte d'efficacité, alors que pour la régression logistique, la TFN induit par un taux de refus équivalent est plus élevée.

Ainsi, l'étude menée répond aux objectifs recherchés : évaluation d'une probabilité de défaillance dans le cadre de la tarification d'un produit d'assurance-crédit et présente plusieurs atouts :

- l'automatisation du process d'octroi et de tarification de la garantie en comparaison des méthodes traditionnelles d'analyse crédit utilisées en assurance-crédit
- une réduction substantielle de la probabilité de défaillance par rapport à la fréquence moyenne empirique des portefeuilles de Pouey International
- une efficacité améliorée avec les forêts aléatoires, légèrement dégradée avec la régression logistique, par rapport aux analyses de Pouey International.

Mais elle présente aussi des limites :

- l'inadaptation à des individus présentant des variables incohérentes ou non renseignées
- le taux de refus élevé nécessaire à une diminution substantielle de la probabilité de défaillance, par rapport à la fréquence empirique générale
- elle ne tient pas compte des éventuelles évolutions de la conjoncture économique pour les secteurs concernés

Ce dernier écueil risque de prendre une importance toute particulière à la suite des conséquences économiques de la crise sanitaire Covid-19 et à l'arrêt total pendant plusieurs mois des activités visées dans cette étude. En effet, la modélisation du risque de défaillance s'est appuyée sur des données des années 2014-2015-2016 durant lesquelles la conjoncture économique était stable. A l'inverse, il faut anticiper une véritable explosion du nombre de défaillances dans ce secteur d'activité, ce qui rendra totalement caduques les modèles établis dans cette étude. Il s'agit là d'une problématique d'un événement extrêmement rare à impact extrêmement élevé, bien supérieur aux capacités de couverture de l'ensemble des établissements d'assurance-crédit. Cette problématique est semblable à celle que connaissent les acteurs intervenant dans les garanties contre les pertes d'exploitation suscitant de vifs débats à l'intérieur et à l'extérieur du monde de l'assurance depuis le début de l'année 2020.

Bibliographie

Publications

- *AON (L. Bollaert 2019) – Marché de l'Assurance-Crédit 2019*
- *R. Eisenbeis (1977) – « Pitfalls in the application of discriminant analysis in business », Journal of Finance*
- *L. Breiman (2001) – Random Forests – Machine Learning volume 45*
- *E. Mays, N. Lynas (2010) – Credit Scoring for Risk Managers*
- *C. Refait-Alexandre (2004) – La prévision de la faillite fondée sur l'analyse financière de l'entreprises : un état des lieux – Economie et Prévision, Minerfi – Direction de la prévision, 162, pp. 129 - 147*

Cours

- *C. Chesneau (2019) – Introduction aux arbres de décision (de type CART)*
- *F. Planchet, A. Miseray (2017) – Tarification IARD – Introduction aux techniques avancées*
- *R. Rakotomalala (2017) – Pratique de la Régression Logistique – Université Lumière Lyon 2*

Site internet

- https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/glm.html#regularization

Annexes

Annexe 1 : extraits du programme sous R

Extrait 1 : calcul des corrélations entre les différentes variables et suppression des variables hautement corrélées

```
      #Correlation entre  les variables predictives ####
            #DTF ####
BCorF=DTF[,VarPredF,with=FALSE]
BCorF=BCorF[!is.na(CATEGORIEENTREPRISE) & !is.na(TYPEEFFECTIF)
                  & !is.na(EFFECTIF) & !is.na(NBETABLISSEMENTS)
                  & !is.na(COUVERTUREIMMOBILISATIONS) & !is.na(VULNERABILITEFINANCIERE)
                  ,]
ChangeDTColumnClass(BCorF,as.numeric,is.oldClass= is.factor)
ChangeDTColumnClass(BCorF,as.numeric,is.oldClass= is.integer)


CorF=cor(BCorF)
z <- as.data.frame(as.table(CorF))
x=subset(z, abs(Freq) > 0.5)
x=x[which(x[,"Var1"]!=x[,"Var2"]),]
y=as.character( unique(x[,"Var1"]) )
B=BCorF[,y,with=FALSE]
LoadPackage("psych")
cor.matrix=corr.test(B, method="spearman")
LoadPackage("corrplot")



corrplot(cor.matrix$r, diag = FALSE, order = "FPC",
         tl.pos = "td", tl.cex = 0.5, method = "color", type = "upper",
         p.mat = cor.matrix$p, insig = "label_sig",
         sig.level = c(.001, .01, .05), pch.cex = .9, pch.col = "white")

corrplot(cor.matrix$r, diag = FALSE, order = "FPC",
         tl.pos = "td", tl.cex = 0.5, method = "color", type = "upper",
         p.mat = cor.matrix$p, insig = "blank", sig.level = .05)

corrplot(cor.matrix$r, diag = FALSE, order = "FPC",
         tl.pos = "td", tl.cex = 0.5, method = "shade", type = "upper",
         p.mat = cor.matrix$p, insig = "blank", sig.level = .05)

corrplot(cor.matrix$r, diag = FALSE, order = "FPC",
         tl.pos = "td", tl.cex = 0.5, method = "square", type = "upper",
         p.mat = cor.matrix$p, insig = "blank", sig.level = .05)

corrplot(cor.matrix$r, diag = FALSE, order = "FPC",
         tl.pos = "td", tl.cex = 0.5, method = "circle", type = "upper",
         p.mat = cor.matrix$p, insig = "blank", sig.level = .05)
```

```
            #Highly cor ####

highlyCorF <- findCorrelation(CorF, cutoff = 0.75, verbose = T)
S1F=names(BCorF)[highlyCorF]
S1F

DTF[,(S1F):=NULL]


highlyCorNF <- findCorrelation(CorNF, cutoff = 0.75, verbose = T)
S1NF=names(BCorNF)[highlyCorNF]
S1NF

DTNF[,(S1NF):=NULL]


highlyCorFU <- findCorrelation(CorFU, cutoff = 0.75, verbose = T)
S1FU=names(BCorFU)[highlyCorFU]
S1FU

DTFU[,(S1FU):=NULL]
```

Extrait 2 : apprentissage de la régression logistique et affichage des résultats

```
                #Creation de base test et training
localH2O <- h2o.init()

NF.hex <- as.h2o(DTNF,destination_frame = "NF.hex")

split <- h2o.splitFrame(data = NF.hex,ratios = 0.75)
Train <- split[[1]]
Test <- split[[2]]



                # Train the model  Classification Defaillance en 2017

VaNF=names(Train)[-which(names(Train)=="Defaillance2017" | names(Train)=="Defaillance2a"
                         | names(Train)=="CODEPOSTAL" | names(Train)=="COMMUNE"
                         | names(Train)=="SIREN")]

NF17.glm <- h2o.glm(x = VaNF, # Vector of predictor variable names
                    y = "Defaillance2017", # Name of response/dependent variable
                    training_frame = Train, # Training data
                    seed = 1561849118991,         # Seed for random numbers
                    family = "binomial",    # Outcome variable
                    lambda_search = TRUE,  # Optimum regularisation lambda
                    alpha = 0.5,           # Elastic net regularisation
                    nfolds = 5 ,            # N-fold cross validation
                    remove_collinear_columns=TRUE,       #Enleve les variables colinéaires
                    compute_p_values = FALSE
)


NF17.glm@model$coefficients

# Print the coefficients table
NF17.glm@model$coefficients_table

h2o.performance(NF17.glm,Test)

pred <- h2o.predict(NF17.glm, Test)

PredictionSiren=as.data.table(pred)
baseTest=as.data.table(Test)
PredictionSiren=cbind(Siren=baseTest[,SIREN],PredictionSiren)

setnames(PredictionSiren,old = colnames(PredictionSiren),
         new = c("Siren","Classification Défaillance","Probabilité de survie","Probabilité de défaillance"))

fwrite(PredictionSiren,file = paste0(DosT,"Siren prediction NF glm 2017.csv"),sep=";",dec = ",")


h2o.varimp_plot(NF17.glm, num_of_features = 10) #Variables ayant les coefs les plus importants
```

Annexe 2 : tables des coefficients en régression logistique

Régression logistique avec données financières, signalétiques et exogènes, variables discrétisées

names	coefficients	standardized_coefficients
Defaillance.1	3,453094076	3,453094076
CATEGORIEENTREPRISE.PME	0,714680947	0,714680947
TRESORERIENETTEQuartile.Cat1	0,47297426	0,47297426
VarCAN_1Quartile.Cat1	0,470506012	0,470506012
VarCAQuartile.Cat1	0,462099835	0,462099835
FRAISFINANCIERSCAQuartile.Cat4	0,440614557	0,440614557
RESULTATNETQuartile.Cat1	0,410693179	0,410693179
PARTFINANCEMENTSTABLEQuartile.Cat1	0,320012824	0,320012824
KP_TOTALBILANQuartile.Cat1	0,318056364	0,318056364
CAFQuartile.Cat1	0,288535599	0,288535599
VULNERABILITEFINANCIEREQuartile.Cat4	0,286172832	0,286172832
VarDstockQuartile.Cat1	0,269476637	0,269476637
VarTNQuartile.Cat1	0,243056029	0,243056029
TxChomT42016Quartile.Cat4	0,205565699	0,205565699
AGECat.4-10	0,182089678	0,182089678
AUTONOMIEFINANCIEREQuartile.Cat1	0,17267303	0,17267303
KP_TOTALBILANQuartile.Cat2	0,170809599	0,170809599
COUVERTUREIMMOBILISATIONSQuartile.Cat1	0,167512613	0,167512613
EBEQuartile.Cat1	0,149597204	0,149597204
TAUXMARGECOMMERCIALEQuartile.Cat4	0,145384993	0,145384993
SalaireMoyenQuartile.Cat1	0,135160276	0,135160276
ENDETTEMENTCAFQuartile.Cat1	0,100838828	0,100838828
CAFQuartile.Cat2	0,096340848	0,096340848
VarDPOQuartile.Cat4	0,088031667	0,088031667
VarKPQuartile.Cat4	0,087730934	0,087730934
VarDPOQuartile.Cat1	0,086941133	0,086941133
EFFECTIFCat.21-100	0,081393679	0,081393679
INDEPENDANCEFINANCIEREQuartile.Cat4	0,080593606	0,080593606
ACTIVITEPRINCIPALE.Commerce de detail	0,064968838	0,064968838
PARTSALAIRESVAQuartile.Cat4	0,061638259	0,061638259
VarKPQuartile.Cat1	0,061448253	0,061448253
VarCAQuartile.Cat2	0,060924415	0,060924415
TAUXMARGECOMMERCIALEQuartile.Cat2	0,058754394	0,058754394
PARTSALAIRESVAQuartile.Cat1	0,047366731	0,047366731

EvolPop5aQuartile.Cat2	0,047157502	0,047157502
TYPEEFFECTIF.1	0,046028933	0,046028933
VENTEFONDS.1	0,033070866	0,033070866
AUTONOMIEFINANCIEREQuartile.Cat4	0,023334856	0,023334856
CHANGTACTIVITE.0	0,022371866	0,022371866
RESULTATNETQuartile.Cat2	0,021188509	0,021188509
RESULTATEXPLOITATIONCAQuartile.Cat4	0,009695291	0,009695291
VarTNQuartile.Cat4	0,008833824	0,008833824
PARTFINANCEMENTSTABLEQuartile.Cat2	0,006451297	0,006451297
CAPITALQuartile.Cat2	0,003120419	0,003120419
TRESORERIENETTEQuartile.Cat4	-0,010495731	-0,010495731
VarCAN_1Quartile.Cat2	-0,01116402	-0,01116402
CREATIONETS.1	-0,012139994	-0,012139994
TxChomT42016Quartile.Cat3	-0,015219075	-0,015219075
LIQUIDITEREDUITEQuartile.Cat3	-0,015620751	-0,015620751
PARTFINANCEMENTSTABLEQuartile.Cat3	-0,020640685	-0,020640685
VarTNQuartile.Cat3	-0,028220686	-0,028220686
TAUXMARGECOMMERCIALEQuartile.Cat3	-0,041249827	-0,041249827
VarKPQuartile.Cat3	-0,06396322	-0,06396322
VarKPQuartile.Cat2	-0,075604789	-0,075604789
ENDETTEMENTCAFQuartile.Cat2	-0,085383069	-0,085383069
CATEGORIEENTREPRISE.ETI	-0,085846857	-0,085846857
SalaireMoyenQuartile.Cat4	-0,086884307	-0,086884307
ACTIVITEPRINCIPALE.Commerce Auto	-0,100421019	-0,100421019
FERMETUREETS.1	-0,100446746	-0,100446746
PARTSALAIRESVAQuartile.Cat3	-0,100551469	-0,100551469
RESULTATNETQuartile.Cat3	-0,100900117	-0,100900117
VarDstockQuartile.Cat3	-0,110382086	-0,110382086
EvolPop5aQuartile.Cat1	-0,111310968	-0,111310968
CHANGTACTIVITE.1	-0,113064488	-0,113064488
RESULTATEXPLOITATIONCAQuartile.Cat2	-0,12179335	-0,12179335
TYPEEFFECTIF.0	-0,125577196	-0,125577196
CAPITALQuartile.Cat1	-0,127798927	-0,127798927
TAUXMARGECOMMERCIALEQuartile.Cat1	-0,134401643	-0,134401643
EFFECTIFCat.0-1	-0,155992957	-0,155992957
CAFQuartile.Cat3	-0,176071671	-0,176071671
EBEQuartile.Cat4	-0,184978015	-0,184978015
CAFQuartile.Cat4	-0,200896105	-0,200896105
VarCAQuartile.Cat4	-0,202990353	-0,202990353
VarTNQuartile.Cat2	-0,214179754	-0,214179754
VarDPOQuartile.Cat2	-0,233883531	-0,233883531

VULNERABILITEFINANCIEREQuartile.Cat2	-0,239685423	-0,239685423
PARTFINANCEMENTSTABLEQuartile.Cat4	-0,28072634	-0,28072634
AUTONOMIEFINANCIEREQuartile.Cat2	-0,295651041	-0,295651041
RESULTATNETQuartile.Cat4	-0,310488445	-0,310488445
VarCAQuartile.Cat3	-0,316647241	-0,316647241
COUVERTUREIMMOBILISATIONSQuartile.Cat3	-0,447836055	-0,447836055
KP_TOTALBILANQuartile.Cat4	-0,696907785	-0,696907785
Intercept	-6,758118914	-6,758118914

Régression logistique sans données financières, variables discrétisées, base de données exhaustive

names	coefficients	standardized_coefficients
Defaillance.1	1,347896607	1,347896607
NatureJuridique.Societe en commandite	0,958799113	0,958799113
AGECat.0-3	0,713566009	0,713566009
CATEGORIEENTREPRISE.PME	0,512183849	0,512183849
NatureJuridique.Societe a responsabilite limitee (SARL)	0,423516468	0,423516468
CHANGTDIRIGEANT.1	0,421575382	0,421575382
CREATIONETS.1	0,38623794	0,38623794
ACTIVITEPRINCIPALE.Restauration	0,349740398	0,349740398
NatureJuridique.Societe par actions simplifiee	0,19344823	0,19344823
EFFECTIFCat.2-5	0,190509101	0,190509101
NBETABLISSEMENTSCat.0-1	0,179037605	0,179037605
FERMETUREETS.1	0,166823807	0,166823807
EFFECTIFCat.0-1	0,116267164	0,116267164
MODIFCAPITAL.1	0,104803286	0,104803286
CHANGTNOM.1	0,086986107	0,086986107
EvolPop5aQuartile.Cat2	0,084079307	0,084079307
CHANGTFJ.1	0,082937709	0,082937709
SalaireMoyenQuartile.Cat1	0,079482519	0,079482519
CHANGTADRESSE.1	0,060181349	0,060181349
SalaireMoyenQuartile.Cat2	0,044371974	0,044371974
CHANGTACTIVITE.1	0,0283284	0,0283284
TxChomT42016Quartile.Cat4	0,027086727	0,027086727
CAPITALQuartile.Cat1	0,022082069	0,022082069
TxChomT42016Quartile.Cat1	0,019510838	0,019510838
CAPITALQuartile.Cat2	0,009539087	0,009539087
EvolPop5aQuartile.Cat3	0,000719458	0,000719458

SalaireMoyenQuartile.Cat3	-0,005822579	-0,005822579
VENTEFONDS.1	-0,017971988	-0,017971988
NBETABLISSEMENTSCat.4-9	-0,019852763	-0,019852763
EvolPop5aQuartile.Cat1	-0,027013545	-0,027013545
TxChomT42016Quartile.Cat3	-0,028319929	-0,028319929
CAPITALQuartile.Cat3	-0,036581613	-0,036581613
MODIFMPROD.0	-0,041339433	-0,041339433
ACTIVITEPRINCIPALE.Commerce de detail	-0,042168341	-0,042168341
CHANGTNOM.0	-0,043607018	-0,043607018
EvolPop5aQuartile.Cat4	-0,069011524	-0,069011524
AGECat.+30	-0,071030099	-0,071030099
TxChomT42016Quartile.Cat2	-0,085339977	-0,085339977
EFFECTIFCat.6-20	-0,101316155	-0,101316155
CHANGTACTIVITE.0	-0,106735137	-0,106735137
CAPITALQuartile.Cat4	-0,116358274	-0,116358274
TYPEEFFECTIF.0	-0,117320612	-0,117320612
SalaireMoyenQuartile.Cat4	-0,152959475	-0,152959475
AGECat.4-10	-0,201765536	-0,201765536
CATEGORIEENTREPRISE.GE	-0,225891879	-0,225891879
EFFECTIFCat.21-100	-0,251194988	-0,251194988
CHANGTADRESSE.0	-0,253242034	-0,253242034
ACTIVITEPRINCIPALE.Hebergement	-0,271222848	-0,271222848
CATEGORIEENTREPRISE.ETI	-0,310111587	-0,310111587
NatureJuridique.Societe en nom collectif	-0,370624969	-0,370624969
NatureJuridique.Association loi 1901 ou assimile	-0,383874819	-0,383874819
NatureJuridique.Entrepreneur individuel	-0,882319305	-0,882319305
Defaillance.0	-1,561616939	-1,561616939
Intercept	-3,310215045	-3,310215045

Annexe 3 : exemple de calcul du seuil optimal selon le score choisi

```
[1] "f1"
[1] "Le seuil f1 est 0.0646940504280584"
   Actual\\Predict Predict 0 Predict 1     Erreur Taux Taux en poucentage
1:        Actual 0     85459      5270 0.05808507   FP          55.820358
2:        Actual 1      5007      4171 0.54554369   FN           5.534676
[1] "f2"
[1] "Le seuil f2 est 0.0273943620474385"
   Actual\\Predict Predict 0 Predict 1    Erreur Taux Taux en poucentage
1:        Actual 0     71149     19580 0.2158075   FP          74.536526
2:        Actual 1      2489      6689 0.2711920   FN           3.380048
[1] "f0point5"
[1] "Le seuil f0point5 est 0.168333065287542"
   Actual\\Predict Predict 0 Predict 1     Erreur Taux Taux en poucentage
1:        Actual 0     88849      1880 0.02072105   FP          39.679190
2:        Actual 1      6320      2858 0.68860318   FN           6.640818
[1] "accuracy"
[1] "Le seuil accuracy est 0.204062207426989"
   Actual\\Predict Predict 0 Predict 1     Erreur Taux Taux en poucentage
1:        Actual 0     89360      1369 0.01508889   FP          36.614068
2:        Actual 1      6808      2370 0.74177381   FN           7.079278
[1] "precision"
[1] "Le seuil precision est 0.337889018906997"
   Actual\\Predict Predict 0 Predict 1     Erreur Taux Taux en poucentage
1:        Actual 0     90429       300 0.00330655   FP          29.154519
2:        Actual 1      8449       729 0.92057093   FN           8.544873
[1] "recall"
[1] "Le seuil recall est 0.000104491093858561"
   Actual\\Predict Predict 0 Predict 1       Erreur Taux Taux en poucentage
1:        Actual 0       140     90589 0.9984569432   FP          90.8014755
2:        Actual 1         1      9177 0.0001089562   FN           0.7092199
[1] "specificity"
[1] "Le seuil specificity est 0.883423069119453"
   Actual\\Predict Predict 0 Predict 1       Erreur Taux Taux en poucentage
1:        Actual 0     90728         1 1.102183e-05   FP         100.000000
2:        Actual 1      9178         0 1.000000e+00   FN           9.186635
[1] "absolute_mcc"
[1] "Le seuil absolute_mcc est 0.13123640197974"
   Actual\\Predict Predict 0 Predict 1     Erreur Taux Taux en poucentage
1:        Actual 0     87806      2923 0.03221682   FP          46.162350
2:        Actual 1      5769      3409 0.62856832   FN           6.165108
[1] "min_per_class_accuracy"
[1] "Le seuil min_per_class_accuracy est 0.0246233621721572"
   Actual\\Predict Predict 0 Predict 1    Erreur Taux Taux en poucentage
1:        Actual 0     68557     22172 0.2443761   FP          76.082630
2:        Actual 1      2208      6970 0.2405753   FN           3.120187
[1] "mean_per_class_accuracy"
[1] "Le seuil mean_per_class_accuracy est 0.0234228296960011"
   Actual\\Predict Predict 0 Predict 1    Erreur Taux Taux en poucentage
1:        Actual 0     67290     23439 0.2583408   FP          76.758580
2:        Actual 1      2081      7097 0.2267379   FN           2.999813
```