Mémoire présenté le :

pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA et l'admission à l'Institut des Actuaires

Par : Aurélien BRUGER

Titre : Modélisation de la cessation d'entreprise et application pour un modèle de résiliation en Assurance Multirisques Commerce

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

Membres présents du jury de l'Institut des Actuaires — Signature

.....

.....

.....

Membres présents du jury de l'ISFA

.....

.....

.....

Entreprise :
Nom : GENERALI France

Signature :

Directeur de mémoire en entreprise :
Nom : Emilie CUZON

Signature :

Invité :
Nom :

Signature :

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

Signature du responsable entreprise

Signature du candidat

Résumé

Les industries collectent de plus en plus de données et sont capables d'en tirer profit au maximum grâce à des algorithmes d'apprentissage automatique.

Par exemple une compagnie d'assurance peut désormais exploiter ces nouvelles données pour améliorer sa rentabilité, cibler ses meilleurs clients et augmenter leur rétention, ou améliorer la rentabilité de son réseau de distribution.

Parmi ces nouvelles données, ce mémoire s'intéresse à celles issues de l'*Open Data*.

Dans un premier temps ce mémoire propose de construire un score de cessation d'activité à partir de données *Open Data* sur chaque entreprise française.

Les agents généraux de Generali France peuvent ainsi utiliser ce nouvel indicateur dans leur prospection de nouveaux clients en privilégiant la prospection des entreprises à faible risque de cessation. En effet lorsqu'une entreprise cesse son activité, tous ses contrats d'assurance peuvent être résiliés et ainsi engendrer une perte de chiffre d'affaires pour l'agent.

Enfin, nous avons constaté que la cessation d'activité représente $20,8\%$ des résiliations des contrats Entreprise & Professionnel de Generali France en 2018.

C'est pourquoi il peut être intéressant d'ajouter le score de cessation comme variable explicative dans un modèle de résiliation d'un produit Entreprise & Professionnel et d'observer si cela améliore son pouvoir prédictif.

Mots-clés : Scoring, *Open Data*, XGBoost, CART, GLM, Forêt aléatoire, Rentabilité, Résiliation

Abstract

Industries are collecting more and more data and are able to benefit from it to the maximum through machine learning algorithms.

For example, an insurance company can now use this new data to improve its profitability, target its customers, target its best customers and increase their retention, or improve the profitability of its distribution network.

Among these new data, this thesis focuses on those from *Open Data*.

As a first step, this thesis proposes to construct a cessation of activity score based on Open Data data on each French company.

Generali France's general agents can use this new indicator in their prospecting of new customers by focusing on prospecting companies with a low risk of cessation. Indeed, when a company ceases its activity, all its insurance contracts may be terminated and thus generate a loss of revenue business for the agent.

Finally, we noted that cessation of activity represents 20.8% of the terminations of Generali France's Enterprise Professional contracts in 2018. This is why it may be interesting to add the cessation of activity score as a predictor variable in a termination model of a Business Professional product and to see if this improves his predictive power.

Keywords : Scoring, *Open Data*, XGBoost, CART, GLM, Random Forest, Profitability, Termination

Remerciements

Tout d'abord je tiens à remercier ma tutrice d'entreprise Emilie CUZON et mon responsable Guillaume VIGNAL pour leur encadrement, leur disponibilité et l'aide qu'ils m'ont apportée tout au long de la réalisation de ce mémoire.

Je remercie également les autres membres de l'équipe *Études Intermédiaires* de Generali France pour les bons moments partagés.

Table des matières

Table des figures

Table des acronymes

ACPR Autorité de Contrôle Prudentiel et de Résolution 62
AUC Area Under the ROC Curve 38
BODACC Bulletin Officiel Des Annonces Civiles et Commerciales 24
CART Classification And Regression Trees 48
ESS Économie sociale et solidaire 17
IML Machine Learning Interpretable 62
INSEE Institut national de la statistique et des études économiques 15
MRC Assurance Multirisques Commerce 12
NAF Nomenclature d'Activité Française 17
PME Petites et Moyennes Entreprises 12
RCS Registre du Commerce et des Sociétés 24
RGPD Règlement Général sur la Protection des Données 62
SARL Société A Responsabilité Limitée 16

Introduction

Le marché de l'assurance des professionnels et des entreprises en France est en plein essor bien qu'il soit aussi très concurrentiel. Il est en effet porté par les perspectives de croissance des 140 000 Petites et Moyennes Entreprises (PME) françaises.

Philippe Protais, directeur de la souscription des risques entreprises chez Generali France, ajoute même à l'Argus de l'assurance « les PME sont une cible privilégiée pour les compagnies comme les réseaux de distributeurs en raison aussi de leur "fort potentiel" de multi-équipements.» [1].

Cependant, en voulant développer son chiffre d'affaires sur les PME, Generali s'expose au risque de cessation d'activité : c'est-à-dire le risque que l'entreprise cesse toute activité et que tous ses contrats d'assurance soient résiliés par la loi.

En 2018, la cessation d'activité est responsable de $20,8\%$ des résiliations sur les produits Generali destinés aux professionnels et entreprises. C'est une part importante du chiffre d'affaires qui est perdu.

C'est pourquoi pour être toujours plus compétitive sur ce marché et protéger son chiffre d'affaires, une compagnie d'assurance peut actionner plusieurs leviers parmi lesquels :

— Piloter plus efficacement la souscription des nouveaux clients professionnels et entreprises en privilégiant les meilleurs risques
— Garder ses clients les plus rentables en portefeuille et limiter leur taux de résiliation

Ce mémoire souhaite contribuer à ces deux enjeux qui sont étroitement liés.

Dans une première partie nous nous intéresserons au contexte de cette étude. D'abord nous présenterons le réseau de distribution de Generali France, puis le cadre juridique des entreprises en France et nous présenterons la notion de cessation d'activité et ses conséquences pour l'assurance. Enfin nous ferons le lien entre la cessation d'activité et la résiliation en assurance à travers l'exemple de l'Assurance Multirisques Commerce (MRC).

Dans une seconde partie, nous allons décrire et modéliser le phénomène de cessation d'activité par un niveau de risque qui sera communiqué aux agents de Generali France pour les aider à démarcher des entreprises peu exposées à ce risque.

Enfin, dans une dernière partie nous mesurerons l'influence de la cessation d'activité sur la résiliation en MRC puis nous essayerons d'améliorer un modèle de résiliation actuel en ajoutant la nouvelle variable du niveau de risque de cessation d'activité.

1. https://www.argusdelassurance.com/assurance-dommages/risques-d-entreprise/marche-des-pme-se-liberer-du-superflu.135634

Première partie

Partie I : Contexte de l'étude

Chapitre 1

Objectifs et présentation de l'étude

1.1 Les réseaux de distribution de Generali France

En 2018 le chiffre d'affaires de Generali France en biens et responsabilité est de 2.7 milliards d'euros. Il est réparti de la façon suivante :

— 60% sur les produits Entreprises et Professionnels
— 32% sur les produits Particuliers
— 8% sur d'autres produits

Cette étude s'inscrit dans le développement du portefeuille Entreprises et Professionnels de Generali France. Generali vend principalement à ses clients Entreprises et Professionnels les produits suivants : Multirisques Commerce, Flottes, RC Pro...

Pour vendre ses produits, Generali s'appuie en grande partie sur son réseau d'intermédiaires composé d'agents généraux et de courtiers.

1.1.1 Les réseaux traditionnels

• Les agents généraux :
Ils sont mandatés par la société qui les emploie. Théoriquement ils ne peuvent pas proposer de produits d'une autre société car ils peuvent en parallèle avoir une affaire de courtage. Les agents généraux sont rémunérés à la commission qui est calculée sur la cotisation nette des contrats qu'ils ont en portefeuille. Le rôle premier d'un agent général est de représenter la société qui l'a mandaté dans le secteur géographique qui lui a été imparti. Ils constituent le groupe le plus important en nombre et en chiffre d'affaires.

• Les courtiers :
Ils sont mandataires de leurs clients. Inscrits au Registre du commerce, ils ne sont liés à aucune société d'assurances et recherche en toute indépendance les produits qui garantissent au mieux les intérêts de leurs clients. Leurs actes n'engagent nullement les assureurs auprès desquels les contrats sont placés (sauf s'ils sont titulaires d'un mandat de gestion). En tant que mandataire de l'assuré, le rôle du courtier est de mettre en rapport toute personne physique ou morale désirant s'assurer avec une société d'assurance en vue de la couverture d'un ou plusieurs risques.

1.1.2 Les autres réseaux

• Les réseaux salariés :
Les réseaux salariés permettent la vente directe de contrats par des assureurs à des clients. Il n'y a alors pas d'intermédiaire entre la compagnie d'assurances et le client. Ces réseaux salariés disposent d'une marge de manœuvre beaucoup plus forte pour faire évoluer leur modèle, dans la mesure où le réseau est intégré.

• Internet :
Internet permet de conclure directement un contrat avec la compagnie d'assurances en ligne. Les compagnies d'assurances investissent dans le développement de plateformes en ligne afin d'offrir de nouveaux services aux clients ou de les améliorer. La souscription d'un contrat d'assurance en ligne permet de souscrire et gérer son contrat en ligne sans jamais avoir besoin d'aller en agence.

Lorsqu'un intermédiaire vend un contrat d'assurance à un nouveau client, il a tout intérêt à le fidéliser en lui proposant des gestes commerciaux ou en le multi-équipant.

1.2 Premier objectif du mémoire

Afin d'améliorer son chiffre d'affaires et sa rentabilité, un intermédiaire doit être vigilant à la souscription d'un nouveau client :

— souscrire un contrat d'assurance représente un coût : diffusion de prospectus, temps de démarchage de l'intermédiaire ou rabais sans conditions
— qualité du risque : certains produits ne sont véritablement rentables pour la compagnie qu'au bout de plusieurs années de cotisation du client. Quelques mesures de qualité de risque sont les suivantes : antécédents du client, position géographique pour le risque inondation ou la viabilité de l'entreprise

Le premier objectif de ce mémoire est de quantifier la viabilité d'une entreprise en construisant un score de cessation d'activité d'entreprise.

Ce nouvel indicateur sera mis à disposition des agents généraux de Generali France. Il leur permettra de piloter leur souscription plus efficacement et de protéger leur chiffre d'affaires : les agents pourront en effet cesser de démarcher des clients avec un risque élevé de cessation d'activité.

1.3 Cadre juridique des entreprises

Il est nécessaire de présenter certains aspects légaux et juridiques des entreprises pour faciliter la lecture du mémoire.

1.3.1 Définition d'une entreprise

L'Institut national de la statistique et des études économiques (INSEE) définit une entreprise de la façon suivante : *L'entreprise est la plus petite combinaison d'unités légales qui constitue une unité organisationnelle de production de biens et de services jouissant d'une certaine autonomie de décision, notamment pour l'affectation de ses ressources courantes.*

Aucune entreprise ne peut s'exempter de l'équilibre entre le niveau de ses revenus et de ses charges. En cas de déficit, celui-ci doit être réduit ou comblé sous peine de non-viabilité et de disparition de l'entreprise à échéance.

Une entreprise est identifiée par un unique identifiant. Il s'agit du numéro Siren composé de neuf chiffres. Il n'est attribué qu'une seule fois et n'est supprimé du répertoire qu'au moment de la disparition de la personne juridique (décès ou cessation de toute activité pour une personne physique, dissolution pour une personne morale).

1.3.2 Statuts juridiques d'une entreprise

En France, une entreprise est nécessairement associée à l'un des deux statuts juridiques suivants : la société ou l'entreprise individuelle. Ces deux statuts se distinguent par leur formalité de constitution, leur régime sociaux et fiscaux, et leur fonctionnement qui est différent.

1.3.2.1 La société morale

L'INSEE définit la société morale de cette façon : *Une société est une entité dotée de la personnalité juridique. Elle est créée dans un but marchand, à savoir, produire des biens ou des services pour le marché, qui peut être une source de profit ou d'autres gains financiers pour son ou ses propriétaires ; elle est la propriété collective de ses actionnaires, qui ont le pouvoir de désigner les administrateurs responsables de sa direction générale.*

1.3.2.2 L'entreprise individuelle

L'INSEE définit l'entreprise individuelle de cette façon : *Une entreprise individuelle est une entreprise qui est la propriété exclusive d'une personne physique. L'entrepreneur exerce son activité sans avoir créé de personne juridique distincte. Les différentes formes d'entreprises individuelles sont : commerçant, artisan, profession libérale, agriculteur. Chaque entreprise individuelle (comme chaque société) est répertoriée dans le répertoire Sirene.* Dans cette étude les auto-entrepreneurs sont inclus dans l'entreprise individuelle.

1.3.3 Natures d'une entreprise

Chaque entreprise possède également une unique nature juridique selon son statut :

- une société morale possède une nature qui décrit sa catégorie juridique (par exemple Société A Responsabilité Limitée (SARL), collectivité territoriale ou association à but non lucratif)
- un entrepreneur individuel possède une nature qui décrit sa catégorie professionnelle (par exemple artisan, commerciale ou libérale)

1.3.4 Entreprises et établissements

Une société morale peut posséder plusieurs établissements, un entrepreneur individuel non. L'INSEE définit un établissement de la façon suivante :

L'établissement est une unité de production géographiquement individualisée, mais juridiquement dépendante de l'entreprise. Un établissement produit des biens ou des services : ce peut être une usine, une boulangerie, un magasin de vêtements, un des hôtels d'une chaîne hôtelière, la « boutique » d'un réparateur de matériel informatique...

Tout établissement possède un unique code Nomenclature d'Activité Française (NAF) qui décrit son activité économique (par exemple Construction, Hébergement et Restauration ou Industrie).

Lorsqu'une entreprise n'exerce pas son activité dans un seul établissement, l'un d'entre eux a le statut d'établissement principal (pour les entreprises individuelles) ou de siège social (pour les sociétés morales).

Dans ce mémoire nous étudierons seulement la cessation d'activité à la maille entreprise et non de l'établissement : un établissement qui cesse son activité n'implique pas que toute l'entreprise est en cessation.

1.3.5 Autres notions sur les entreprises utilisées dans l'étude

Une entreprise peut appartenir au champ de l'économie sociale et solidaire. La loi n° 2014 – 856 du 31 juillet 2014 définit les principes pour qu'une entreprise adhère à l' Économie sociale et solidaire (ESS) :

- poursuivre un but social autre que le seul partage des bénéfices
- une lucrativité encadrée (notamment des bénéfices majoritairement consacrés au maintien et au développement de l'activité)
- une gouvernance démocratique et participative

1.4 La cessation d'entreprise

L'INSEE définit la cessation d'entreprise de la façon suivante : *La cessation d'entreprise correspond à l'arrêt total de l'activité économique de l'entreprise.*

Pour les entreprises individuelles, cela correspond au dépôt de la déclaration de la disparition de l'entreprise individuelle. Pour les personnes physiques, cela correspond soit à la prise en compte de la déclaration de cessation d'activité déposée par l'exploitant soit au décès de l'exploitant conformément à la réglementation.

Les principales raisons qui amènent une entreprise à cesser son activité sont parmi les suivantes : cessation de paiement (incapacité pour l'entreprise de faire face à son passif exigible à court terme par ses actifs immédiatement disponibles), départs à la retraite, décès du dirigeant ou autres difficultés économiques.

Généralement les auto-entrepreneurs cessent davantage leur activité pour les raisons suivantes :

- **le développement de l'activité :** dès que l'activité d'un auto-entrepreneur est bien établie, nombre d'entre eux choisissent de poursuivre leur activité en société
- **le changement d'activité :** pour changer d'activité, il est nécessaire de fermer l'auto-entreprise. Ensuite, l'entrepreneur peut choisir de débuter sa nouvelle activité en créant une nouvelle auto-entreprise ou en créant une société

— **la volonté de travailler en tant que salarié :** certains souhaitent se tourner vers le salariat pour bénéficier d'un cadre plus sécurisant
— **le dépassement des plafonds auto-entrepreneur :** pour être auto-entrepreneur, il faut respecter des seuils de chiffre d'affaires annuel (170 000 euros pour la vente de marchandises et 70 000 euros pour les prestations de services). Au-delà, l'entrepreneur bascule d'office sous le statut de l'entreprise individuelle

Quelle qu'en soit la raison, la cessation d'activité entraîne des obligations légales et fiscales. Après établissement du bilan et du compte de résultat, le procès-verbal de dissolution doit être déposé au greffe du tribunal de commerce. Les bénéfices réalisés depuis le dernier exercice seront quant à eux imposés. Pour les salariés, la cessation d'activité constitue un motif de licenciement économique.

1.5 Lien entre la cessation d'entreprise et la résiliation

Dans un premier temps, lorsqu'une entreprise cesse son activité tous les contrats d'assurance qu'elle a souscrit peuvent être résiliés.

En effet, un contrat d'assurance peut être transféré si la société est vendue à un autre entrepreneur, sinon le contrat est résilié à la date de cessation : les primes futures sont perdues.

Dans un second temps, étudions la répartition des motifs de résiliations en 2018 sur la gamme de produits Entreprises & Professionnels de Generali France.

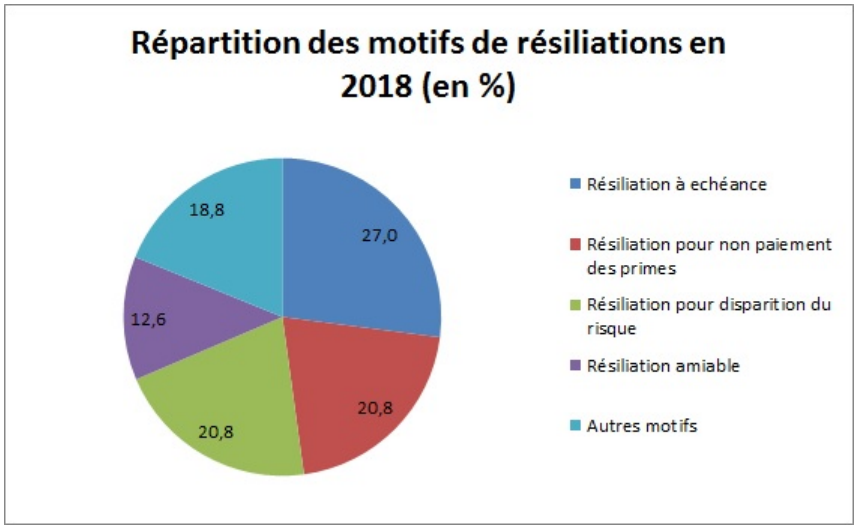

FIGURE 1.1 – Motif de résiliations des produits Entreprise & Professionnel en 2018

Ce graphique nous indique que $20,8\%$ des résiliations de contrats Entreprise & Professionnel sont liés à une "disparition du risque" en 2018, c'est-à-dire que le client a cessé son activité et résilié son contrat.

Il est donc naturel de conclure que la cessation d'activité est un motif important de résiliation des produits Entreprise & Professionnel.

C'est pourquoi une fois que aurons calibré un score de cessation sur les entreprises de la base Sirene©, nous étudierons s'il permet d'améliorer un modèle de résiliation sur

un produit Entreprise & Professionnel. Nous avons alors choisi le produit Assurance Multirisques Commerce (MRC).

1.6 L'Assurance Multirisques Commerce (MRC)

Generali distribue deux produits dans sa branche MRC : *100% Pro Artisans - Commerçants* à destination des artisans, commerçants et petits fabriquants et *100% Pro services* à destination des activités de services et professions libérales.

Ces produits peuvent couvrir les garanties suivantes selon les options choisies par le client :

- **Garanties dommages aux biens**
  - Incendie et vandalisme
  - Dégâts des eaux
  - Vol
  - Bris
- **Garanties responsabilité civile professionnelle**
- **Garanties protection juridique.** Protection professionnelle et commerciale incluant la gestion des litiges liés :
  - Aux locaux professionnels
  - Au quotidien commercial
  - Aux relations employeurs / employés
  - A l'administration et à l'Urssaf
  - A l'administration fiscale

Quelques chiffres sur la MRC pour Generali France

En 2017 Generali occupe la 6$^{\text{ème}}$ position du marché français de la MRC avec 6.7% de part de marché et un chiffre d'affaires de 120 millions d'euros.

1.7 La résiliation

Aujourd'hui la rentabilité des compagnies d'assurance provient en grande partie des contrats déjà en portefeuille. C'est pourquoi une compagnie d'assurance a intérêt à cibler des populations à duration forte tout en ayant une forte rétention de son portefeuille.

Quelques problématiques apparaissent donc :

- Quels sont les principaux leviers qui permettent de garder un client en portefeuille ?
- Quel est le profil des assurés qui ont tendance à résilier ?

Nous avons choisi de modéliser la résiliation en MRC de la façon suivante : nous observons les contrats MRC actifs dans le portefeuille à la date T_0 = 01/07/2017 puis nous regardons à la date $T_0 + 24$ mois = 01/07/2019 les contrats qui ont été résiliés dans cette période.

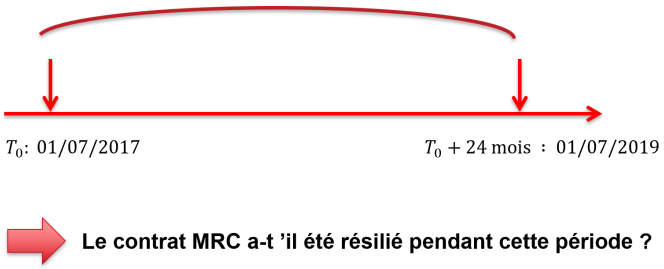

FIGURE 1.2 – Période de résiliation observée

Le périmètre a été choisi de sorte à capturer au mieux l'effet du score de cessation d'entreprise qui est calculé sur des données à juillet 2019.

Pour conclure si le score de cessation améliore ou non ce modèle de résiliation, nous procédons de la façon suivante : nous modélisons la résiliation en MRC avec deux modèles, un modèle avec le score de cessation en variable explicative et un modèle sans, puis nous les comparerons selon certains critères de performance.

1.8 Conclusion : Objectif du mémoire

Pour conclure, l'objectif principal du mémoire est de calibrer un score de cessation d'activité sur les entreprises d'un périmètre défini et de mettre ce score à disposition des agents généraux de Generali France pour les aider à piloter leur souscription sur la branche Entreprise & Professionnel.

Dans un second temps, nous étudierons si ce score peut aussi être utilisé pour améliorer un modèle de résiliation d'un produit Entreprise & Professionnel commercialisé par Generali France.

En effet, nous avons observé que pour cette gamme de produit les résiliations de contrats pour "disparition du risque" représentent plus de 20% des résiliations en 2018.

Ainsi nous avons choisi le produit de la branche Assurance Multirisques Commerce (MRC) où nous modéliserons l'acte de résiliation à 2 ans, et ce avec et sans le score de cessation pour mesurer son influence.

Deuxième partie

Partie II : Modélisation de la cessation d'activité

Chapitre 2

Analyse et préparation des données

Dans cette partie nous présenterons les données utilisées, le périmètre de l'étude puis la modélisation du risque de cessation d'activité ainsi que quelques études descriptives.

2.1 Notions sur l'Open Data

2.1.1 Définition

Les *Open Data*, ou données ouvertes, sont des données auxquelles l'accès est totalement public et libre de droit, au même titre que l'exploitation et la réutilisation : tout le monde peut utiliser ou partager ces données librement et gratuitement.

Son origine remonte à l'année 1966 lorsque les États-Unis instaure la loi pour la liberté d'information (ou *Freedom of Information Act*) qui dessine les prémisses de l'Open Data : les agences fédérales sont obligées de transmettre leurs documents à quiconque en fait la demande, quelle que soit sa nationalité.

En 2005, l'Open Knowledge Foundation donne 3 critères essentiels que l'Open Data doit respecter :

- **Disponibilité et accès :** les données doivent être pleinement accessibles et de préférence pouvoir se télécharger sur internet. Les formats libres de fichiers comme le .txt ou le .csv sont privilégiés par rapport aux formats propriétaires comme les fichiers Excel (.xls ou .xslx).
- **Réutilisation et redistribution :** les données doivent être fournies sous des conditions permettant la réutilisation, la redistribution et l'enrichissement avec d'autres sources de données.
- **Participation universelle :** tout le monde doit être en mesure d'utiliser, de réutiliser et de redistribuer les données. Aucune discrimination concernant les fins d'utilisation, ou contre des personnes ou des groupes n'est possible. Par exemple, des restrictions d'usage à certains secteurs ne sont pas compatibles avec l'Open Data.

2.1.2 L'Open Data en France

Le 5 novembre 2011, la mission Etalab placée sous l'autorité du Premier ministre créé la plate-forme *data.gouv.fr* qui permet un accès libre et une réutilisation gratuite d'un grand nombre de données publiques. Ainsi Etalab mentionne que *"l'ouverture des données d'intérêt public vise à encourager la réutilisation des données au-delà de leur utilisation première par l'administration."*

Dans ce mémoire, les données Sirene et du Registre du Commerce et des Sociétés présentées dans la partie suivantes ont été extraites sur *data.gouv.fr*.

2.2 Données utilisées

2.2.1 Données du répertoire Sirene

La base Sirene rassemble les informations économiques et juridiques sur 28 millions d'établissements actifs et fermés en France appartenant à tous les secteurs d'activité. Les entreprises étrangères qui ont une représentation ou une activité en France y sont également répertoriées.

Elle est mise à disposition librement et gratuitement par l'INSEE et comprend toutes les entreprises identifiées par un numéro Siren, identifiant unique d'une entreprise, d'un organisme ou d'une association. De nombreux organismes déclarent à l'INSEE les immatriculations, radiations et modifications du répertoire.

La base Sirene rassemble des données attachées soit à l'entreprise soit à l'établissement. Voici quelques exemples de variable dans la base Sirene :

Données d'identification pour l'entreprise et l'établissement

Pour l'entreprise :	Pour l'établissement :
Le numéro Siren	Le numéro Siret
Le nom, les prénoms pour les entrepreneurs individuels	Le statut de siège ou non
Le sigle, la raison sociale et la dénomination pour une personne morale	Les enseignes
La catégorie juridique	L'adresse complète sous forme d'éléments d'adresse

Données structurelles pour l'entreprise et l'établissement

Pour l'entreprise :	Pour l'établissement :
Le code d'activité principale (APE) attribué par l'Insee en référence à la nomenclature d'activités française (NAF rév. 2, 2008)	Le code d'activité principale (APE) attribué par l'Insee en référence à la nomenclature d'activités française (NAF rév. 2, 2008)
L'importance de l'effectif salarié de l'entreprise (tranches)	L'importance de l'effectif salarié pour chaque établissement (tranches)

Autres données pour l'entreprise et l'établissement

Pour l'entreprise :	Pour l'établissement :
La date de création	La date de création
Variable de localisation géographique du siège de l'entreprise : la commune	Variable de localisation géographique de l'établissement : la commune
La date du dernier traitement d'une mise à jour des données de l'entreprise dans le répertoire Sirene	La date du dernier traitement d'une mise à jour des données de l'établissement dans le répertoire Sirene

2.2.2 Données du Registre du Commerce et des Sociétés

Le Registre du Commerce et des Sociétés (RCS) est le service tenu par le greffe du Tribunal de commerce qui traite les dossiers de la compétence du Tribunal du commerce. Le RCS collecte des données de 5,5 millions d'entreprises qui exercent une activité commerciale, soit 80% des entreprises françaises.

Dans ce mémoire nous exploitons deux sources de données *Open Data* mises à disposition par le RCS :

- La base de données Chiffres Clés 2016-2018
- La base de données des Procédures Collectives publiées au BODACC

• Chiffres Clés 2016-2018

Toutes les personnes physiques ou morales inscrites au Registre du Commerce et des Sociétés ont l'obligation de déposer leurs comptes annuels auprès du greffe du Tribunal de commerce.

Les entrepreneurs individuels en sont exemptés : ils n'ont qu'à produire un journal des recettes et un registre des achats. Malgré leur exemption, s'ils décident de produire leur compte annuel, ils peuvent bénéficier d'une demande de confidentialité.

Ces comptes annuels sont composés de 3 documents :

- **le bilan :** décrit l'actif et le passif dont dispose la société à la clôture de l'exercice
- **le compte de résultat :** récapitule les produits et les charges de l'exercice ainsi que le résultat de leur différence : il s'agit du *résultat de l'exercice*
- **l'annexe :** complète et commente les données figurant au bilan et au compte de résultat

Une entreprise se doit de produire et de déposer son compte annuel auprès du greffe de tribunal dont elle relève dans les 6 mois qui suivent la clôture de l'exercice comptable.

C'est ainsi que le greffe du Tribunal de commerce met à disposition en Open Data les chiffres clés de certaines sociétés inscrites au RCS ayant déposé leurs comptes annuels pour les exercices 2016, 2017 et 2018. On retrouve ainsi pour ces entreprises les informations suivantes sur les exercices de 2016 à 2018 :

- le chiffre d'affaires de l'exercice
- le résultat de l'exercice
- l'effectif de l'entreprise sur l'exercice

• Procédures collectives publiées au BODACC

Dans le cadre de sa mission de service public de la transparence économique et financière, le Bulletin Officiel Des Annonces Civiles et Commerciales (BODACC) édité par la Direction de l'information légale et administrative assure la publication des actes enregistrés au RCS comme par exemple :

- les ventes et cessions de sociétés
- les créations d'établissements
- l'immatriculation de sociétés
- les procédures collectives

La procédure collective est une procédure de redressement ou de liquidation judiciaire organisant le paiement des créances d'une entreprise en cessation de paiement.

L'entreprise concernée par la procédure collective se voit alors prescrire par le tribunal du Commerce :

- **un redressement judiciaire**, si l'entreprise est en état de cessation des paiements. Elle vise à permettre la poursuite de l'activité, le maintien de l'emploi et l'apurement du passif, en procédant à une réorganisation de l'entreprise dans le cadre d'un plan arrêté par le Tribunal.
- **une liquidation judiciaire**, si l'entreprise est en état de cessation des paiements et que l'activité a cessé ou que le redressement apparaît manifestement impossible. L'entreprise se déclare alors en cessation d'activité.

2.2.3 Zonage en aires urbaines de l'INSEE

En 2010, l'INSEE a créé un zonage en aires urbaines en France pour décrire l'influence des villes sur l'ensemble du territoire. Ce découpage est fondé sur l'identification de pôles, unités urbaines concentrant au moins 1 500 emplois.

2.2.4 Réglementation autour des données utilisées

Les données Sirene et du Registre du Commerce et des Sociétés sont soumises à la « Licence Ouverte / Open Licence » [1] conçue par Etalab. Cette licence permet par exemple la réutilisation des informations en autorisant la reproduction, la redistribution, l'adaptation et l'exploitation commerciale des données sous réserve de mentionner sa provenance.

Ainsi les sources des jeux de données sont mentionnées dans l'annexe 1.

2.2.5 Construction de la variable cible

Dans cette étude de la cessation d'entreprise, la variable à expliquer est la variable binaire $DEFAUT$ qui indique si l'entreprise est active ($DEFAUT = 0$) ou si l'entreprise a cessé son activité ($DEFAUT = 1$) à la date d'extraction de la base Sirene, soit le 01/07/2019 dans ce mémoire.

La variable cible $DEFAUT$ est construite à partir de la variable *etatAdministratifUniteLegale* dans la base Sirene qui vaut "Cessée" lorsque l'entreprise est en cessation administrative.

2.3 Périmètre de l'étude

L'étude concerne les cessations d'activité décidées par les 134 tribunaux de commerce du 1er janvier 2008 au 1er juillet 2019, date d'extraction de la base Sirene : nous retirons donc les entreprises qui ont cessé leur activité avant le 1er janvier 2008.

De plus, les entreprises du secteur agricole, les associations et les professions libérales qui relèvent des tribunaux de grande instance n'entrent pas dans le champ de cette étude. En effet, contrairement aux tribunaux de commerce, les tribunaux de grande instance ne fournissent pas le numéro Siren des entreprises. On ne peut pas

1. lien vers la licence : `https://www.etalab.gouv.fr/wp-content/uploads/2017/04/ETALAB-Licence-Ouverte-v2.0.pdf`

donc pas enrichir les données du répertoire Sirene avec les procédures collectives des entreprises qui ne relèvent pas des tribunaux de commerce.

Les entreprises relevant de la compétence des tribunaux de commerce sont déterminées par les catégories juridiques suivantes :

1 - Entreprises individuelles
- 11 - Artisan-commerçant
- 12 - Commerçant
- 13 - Artisan

3 - Personne morale de droit étranger
- 31 - Personne morale de droit étranger, immatriculée au RCS

5 - Société commerciales (toutes les catégories)

6 - Autre personne morale immatriculée au RCS
- 62 - Groupement d'intérêt économique
- 69 - Autre personne morale de droit privé inscrite au RCS

9 - Groupement de droit privé
- 99 - Autre personne morale de droit privé

Il est maintenant intéressant de comparer le périmètre d'étude avec le portefeuille client entreprise de Generali. Nous avons ainsi croisé la liste des clients entreprises de Generali avec la base Sirene, puis nous l'avons comparé avec notre base d'étude.

Par exemple étudions la répartition des catégories juridiques dans la base d'étude et dans la base client de Generali.

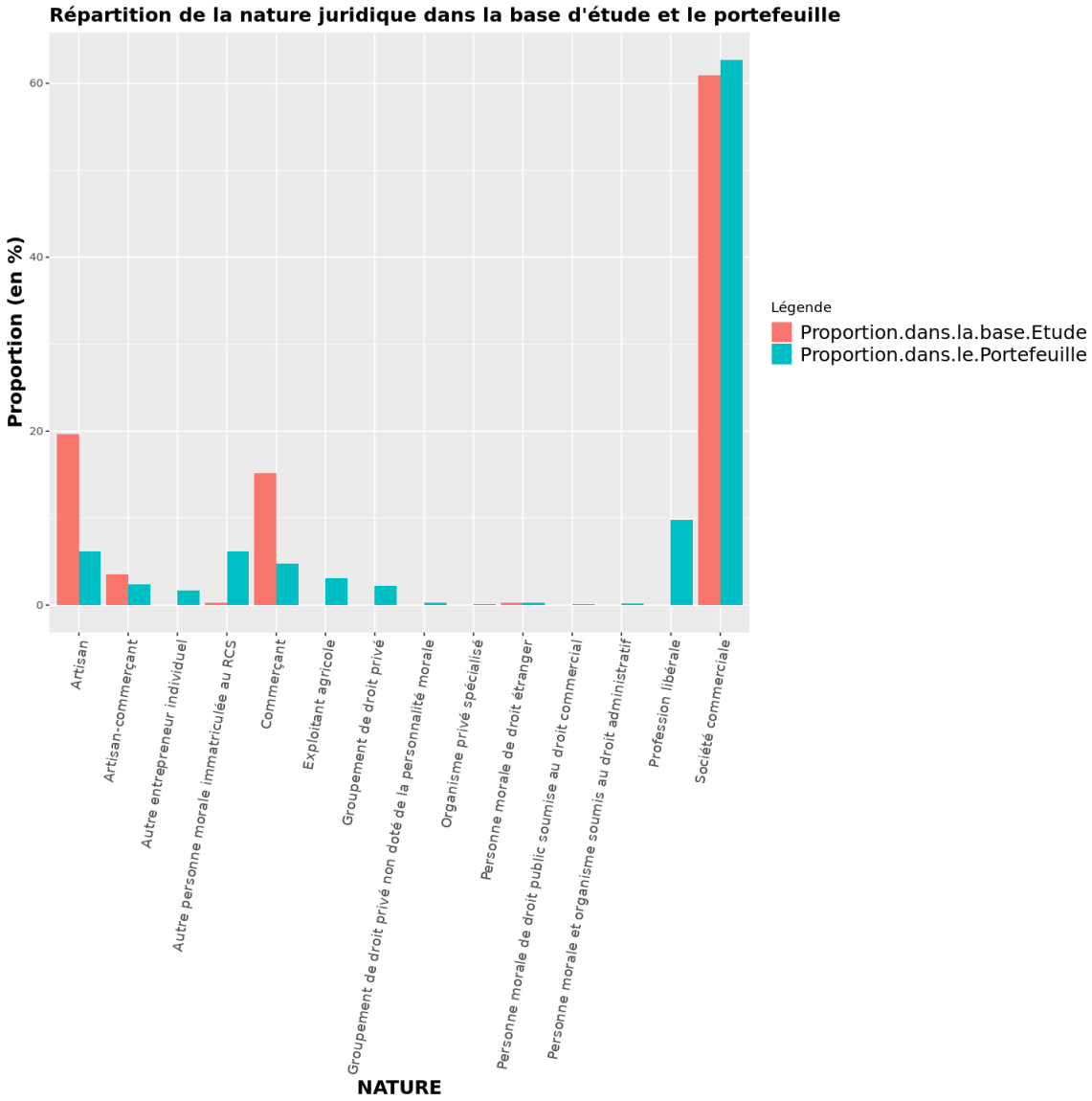

FIGURE 2.1 – Comparaison du portefeuille de Generali France et de la base d'étude par nature juridique

Bien que les professions libérables représentent environ 10% du portefeuille de Generali France, nous les avons exclus du périmètre de modélisation du score de cessation d'activité car nous n'avons pas pu enrichir leurs données.

Toutefois le périmètre établi capte 87% du portefeuille de Generali France.

Ainsi nous validons notre périmètre : nous allons calibrer notre score de cessation d'activité sur les entreprises relevant des tribunaux de commerce actives au 1er janvier 2008.

2.4 Retraitement de la base

Maintenant que le périmètre est établi, il est nécessaire de retraiter la base de données avant la modélisation.

Gestion des données manquantes

Nous identifions plusieurs catégories de données manquantes dans notre étude :

— Les données manquantes parce que non renseignées dans le répertoire Sirene : nous décidons de les remplacer par la médiane, la modalité la plus fréquente ou par "Inconnu".
— Les données manquantes car non observées : par exemple si une entreprise n'a été visée par aucune procédure collective. Nous décidons de les remplacer par "Absent".
— Les données manquantes car confidentielles : par exemple les entreprises individuelles ne sont pas concernées par l'obligation de dépôt des comptes annuels au greffe. De plus les petites sociétés tenues de déposer leurs comptes annuels pouvent demander que ces derniers restent confidentiels en effectuant une déclaration de confidentialité. Nous décidons de les remplacer par "Confidentiel".

Autres retraitements

— Les variables qualitatives avec de nombreuses modalités sont supprimées comme le numéro siren, ou la ville et le nom de l'établissement.
— Certaines variables continues sont regroupées en variable qualitative comme l'âge de l'entreprise ou le nombre d'employés.

2.5 Analyse descriptive de la cessation d'activité

Avant de modéliser la cessation d'activité, il est important d'effectuer une étude descriptive pour mieux comprendre le phénomène et l'impact de certaines variables.

La base finale contient 7 887 087 entreprises actives au 1er janvier 2008 pour 1 834 072 cession d'activités depuis, soit un taux de cessation $23,2\%$.

2.5.1 La cessation d'activité depuis 2008

Dans un premier temps étudions le taux de cessation moyen annuel entre 2008 et 2019. Pour une année N, le taux de cessation rapporte le nombre d'entreprises qui cessent leur activité au cours de l'année N au nombre d'entreprises présentes dans la base de données au premier janvier de l'année N.

$$\textit{Taux de cessation de l'année } N = \frac{\textit{Nombre d'entreprises en cessation l'année N hors création d'entreprise}}{\textit{Nombre d'entreprises présentes au 01/01/N}}$$

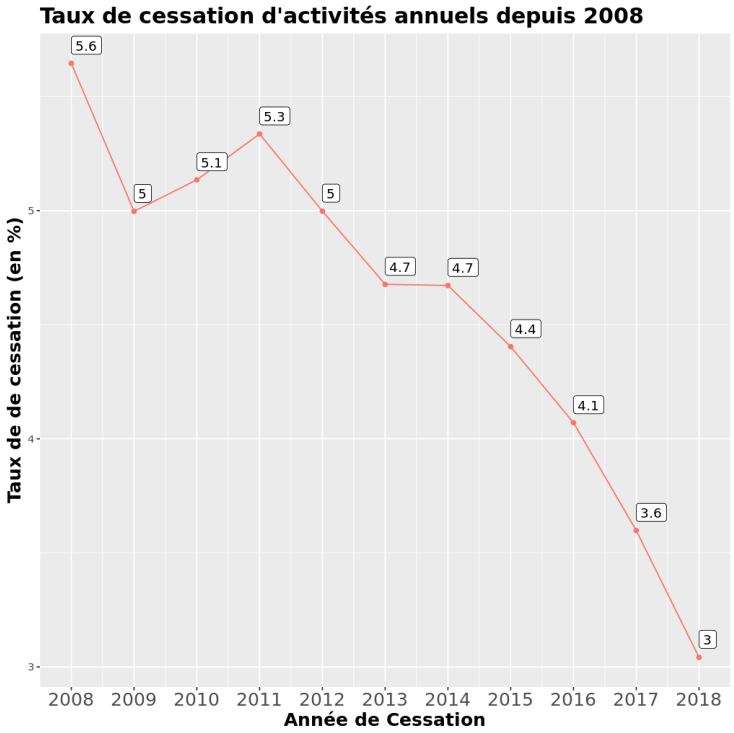

FIGURE 2.2 – Taux de cessation d'activités annuels entre 2008 et 2018

Ce graphique nous indique que la cessation d'activité en France à tendance à diminuer depuis 2008, avec une forte baisse observée depuis 2014.

Cette tendance peut s'expliquer par la hausse du nombre de créations d'entreprises ainsi que la baisse du nombre de défaillances (procédures de redressement judiciaire contre les entreprises) depuis 2014 :

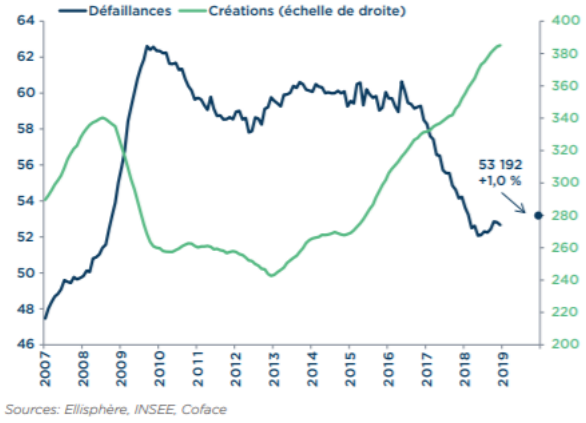

FIGURE 2.3 – Défaillances et créations d'entreprises depuis 2007 (en milliers, annuel), (issu de l'article [6])

2.5.2 Analyse univariée

Pour mieux comprendre le phénomène de cessation d'activité, nous allons maintenant présenter quelques graphiques avec la distribution du taux moyen de cessation d'activité pour quelques variables explicatives.

Par nature juridique

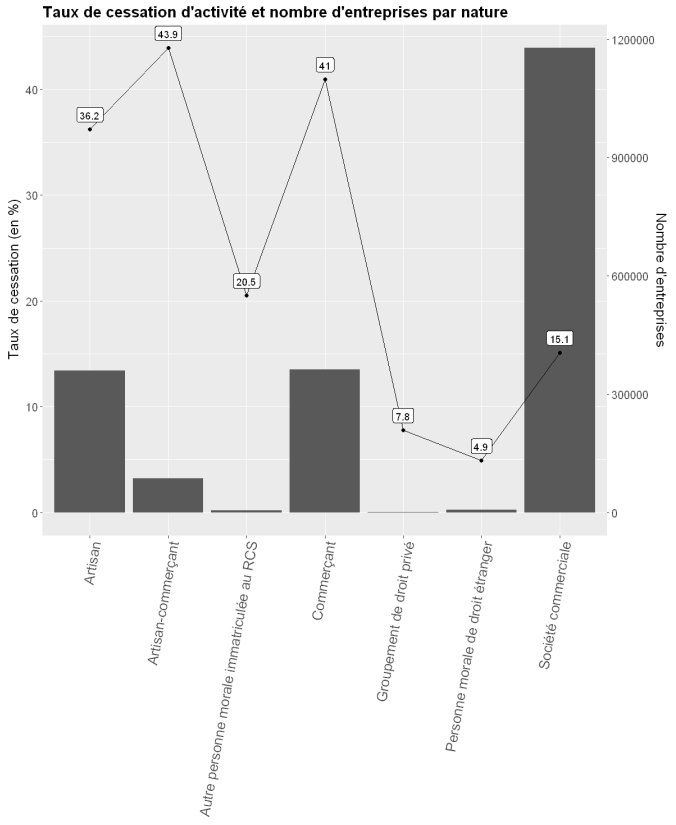

FIGURE 2.4 – Taux de cessation moyen par nature juridique

D'après ce graphique nous observons que la cessation d'activité semble fortement liée à la nature juridique de l'entreprise : les artisans, commerçants et artisans-commerçants cessent davantage leur activité que les sociétés commerciales.

En effet $43,9\%$ des artisans-commerçants ont cessé leur activité depuis 2008 contre seulement $15,1\%$ des sociétés commerciales.

Cela s'explique par le fait que les auto-entrepreneurs ont davantage tendance à cesser leur activité comme mentionné dans le paragraphe 1.4.

Par secteur d'activité

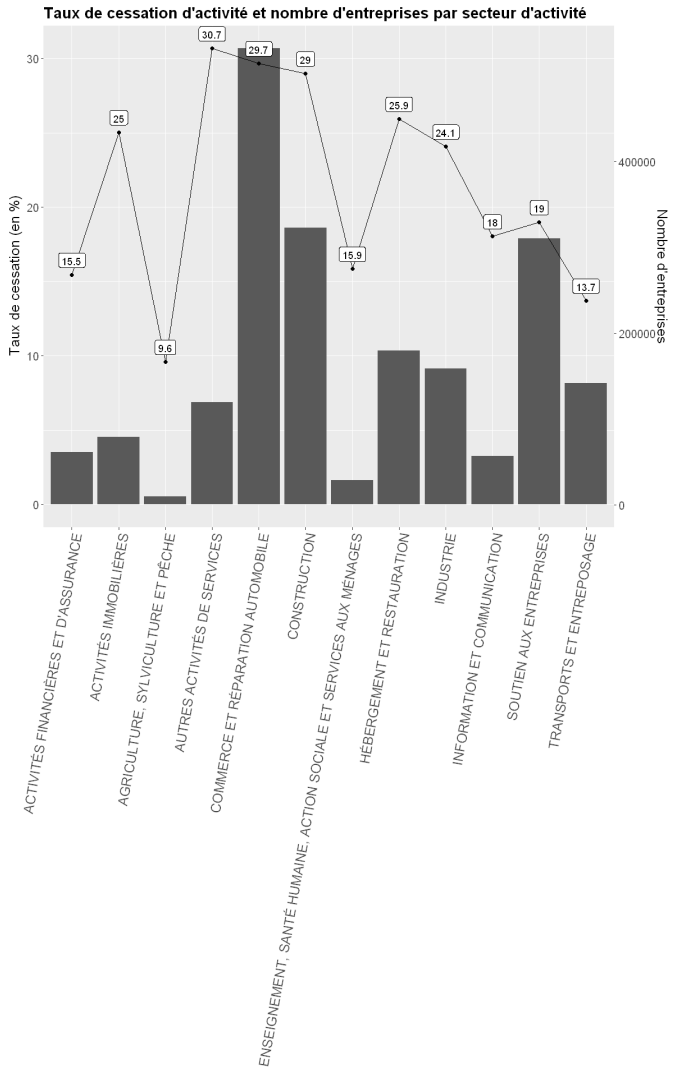

FIGURE 2.5 – Taux de cessation moyen par secteur d'activité

Quelques pistes pour expliquer les taux moyen de cessation élevés de certains secteurs d'activité :

— la *CONSTRUCTION* accuse une baisse importante des permis de construire (- 10, 2% entre les troisièmes trimestres 2017 et 2018) [2]
— les *AUTRES ACTIVITÉS DE SERVICE* peuvent être affectés par le ralentissement de la consommation des ménages (+ 0, 9% en 2018, après + 1, 4% en 2017 et + 1, 8% en 2016) [3]
— le *COMMERCE ET RÉPARATION AUTOMOBILE* souffre d'une baisse du nombre d'immatriculations de voitures particulières neuves (- 8, 5% entre septembre 2017 et septembre 2018) [4]

2. https://www.batiactu.com/edito/nombre-permis-construire-baisse-102--au-troisieme-trimestre-54468.php
3. https://www.insee.fr/fr/statistiques/4168956
4. https://www.autoplus.fr/actualite/Marche-Auto-Chiffres-Ventes-Voitures-particulieres-neuves-Immatriculation-Septembre-2018-1531639.html

Par ancienneté de l'entreprise

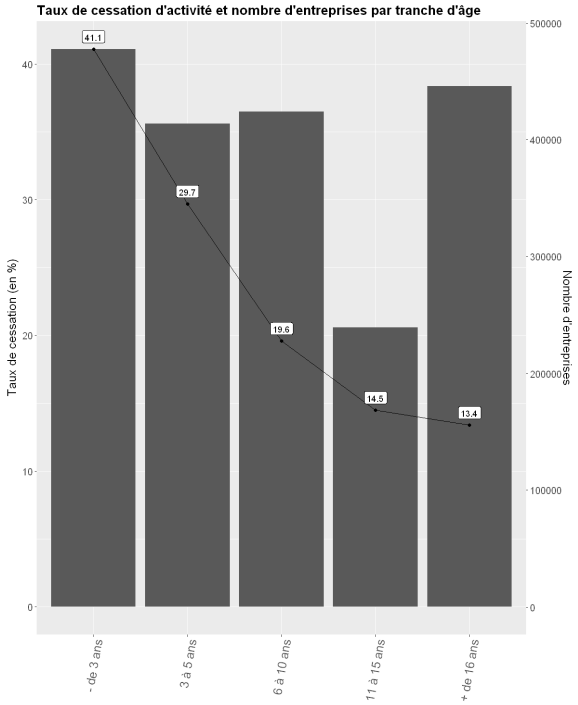

FIGURE 2.6 – Taux de cessation moyen par ancienneté d'entreprise

L'âge de l'entreprise est également très lié à la cessation d'activité. En effet le taux moyen de cessation est de $41,1\%$ parmi les entreprises qui ont moins de 3 ans d'ancienneté, puis il diminue avec l'ancienneté pour atteindre $13,4\%$ parmi les entreprises qui ont plus de 16 ans.

2.5.3 Étude des liaisons

L'étude des liaisons permet de faire une première sélection de variables en gardant uniquement des variables peu corrélées entre elles et en supprimant les variables qui sont corrélées à moins de 1% de la variable cible.

Vu que les variables candidates sont soit qualitatives soit quantitatives discrétisées nous utilisons le V de Cramer comme mesure d'association entre elles. Nous utilisons la même mesure d'association entre les variables candidates et la variable à expliquer car cette dernière est aussi qualitative.

Le V de Cramer se calcule ainsi :

$$V = \sqrt{\frac{\chi^2}{\chi^2_{max}}}$$

avec

$$\chi^2_{max} = \text{effectif} \times [\min(\text{nombre de lignes}, \text{nombre de colonnes}) - 1]$$

Le V de Cramer varie entre 0 (liaison nulle) et 1 (liaison maximale), ce qui évalue l'intensité de la liaison entre les deux variables.

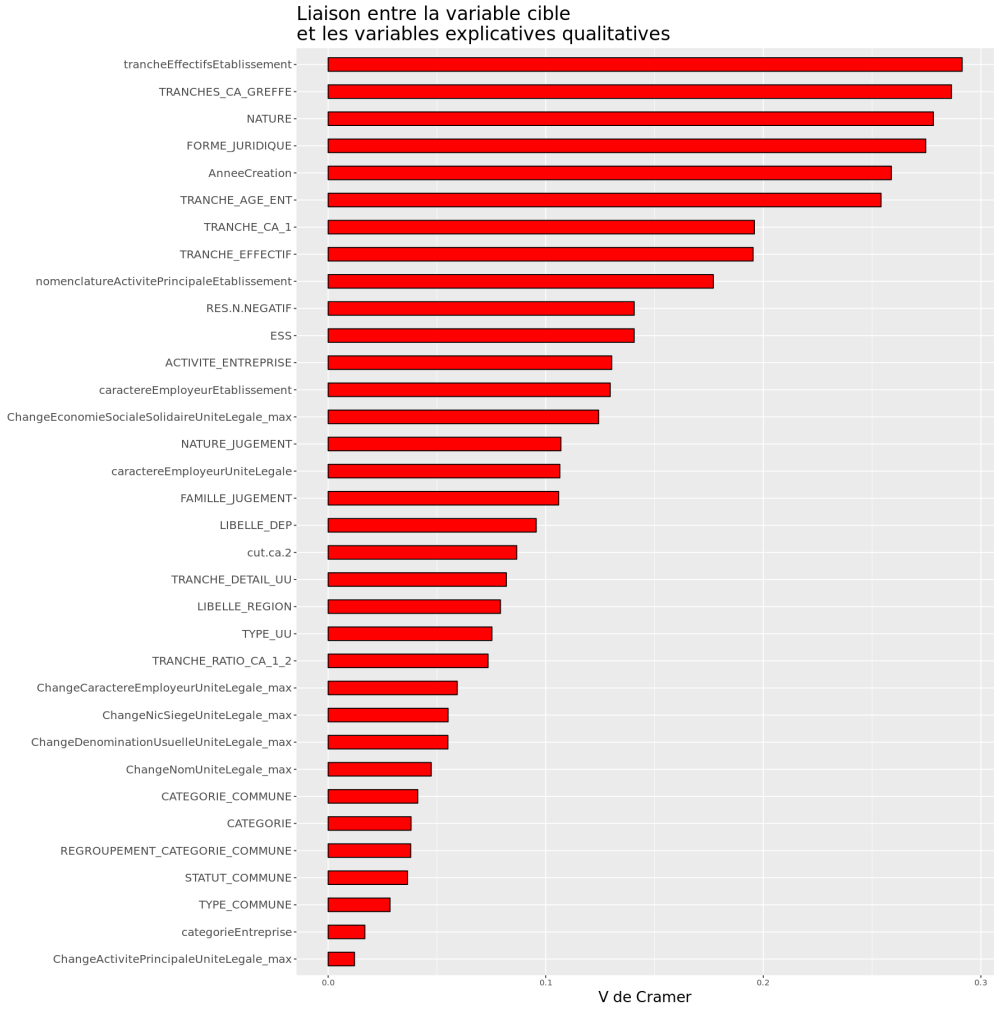

FIGURE 2.7 – V de Cramer entre la variable cible et les variables explicatives

Liaisons entre la variable cible et les variables candidates

Les variables candidates les plus liées à la cessation d'activité d'une entreprise sont sa tranche d'effectif, sa tranche d'âge, sa nature juridique et si elle appartient à l'économie sociale et solidaire.

Nous décidons de supprimer les variables dont le V de Cramer avec la variable cible est inférieure à 0.01 car elle n'apportent pas assez d'informations.

Liaisons entre les variables candidates

Il est important de détecter les colinéarités entre les variables candidates pour améliorer la qualité prédictive des modèles type GLM.

FIGURE 2.8 – V de Cramer entre les variables explicatives

La plupart des variables très liées entre elles représentent le même phénomène à des mailles différentes, par exemple les variables du zonage de l'INSEE. Nous décidons de garder une seule variable par phénomène.

2.6 Variables retenues pour modéliser la cessation d'activité

Après avoir collecté un nombre important de variables issues de l'Open Data, l'analyse des liaisons a permis d'isoler les variables retenues pour modéliser la cessation d'activité.

Ainsi les variables candidates retenues :

— ont une liaison non négligeable avec la variable cible

— ne sont pas trop corrélées entre elles pour ne pas avoir de redondance d'informations

Le tableau ci-dessous résume les variables retenues :

Variable	Description
TRANCHE_EFFECTIF	Tranche du nombre d'employé de l'entreprise
CATEGORIE	Catégorie de l'entreprise selon l'effectif et le chiffre d'affaires
caractereEmployeurEtablissement	Est-ce que l'établissement siège a des employés ?
NATURE	Nature juridique de l'entreprise
LIBELLE_REGION	Région de l'établissement siège
CATEGORIE_COMMUNE	Catégorie de la commune selon le zonage urbain de l'INSEE
ChangeNicSiegeUniteLegale_max	Est-ce que l'établissement siège a déjà changé son code NIC ?
ChangeActivitePrincipaleUnite Legale_max	Est-ce que l'entreprise a déjà changé de secteur d'activité ?
ChangeCaractereEmployeurUniteLegale_max	Est-ce que l'entreprise a déjà changé de caractère employeur ?
ChangeNom	Est-ce que l'entreprise a déjà changé de nom ?
TRANCHE_AGE_ENT	Tranche d'âge de l'entreprise
ESS	Est-ce que la société morale appartient à l'économie sociale et solidaire ?
ACTIVITE_ENTREPRISE	Secteur d'activité de l'entreprise
FAMILLE_JUGEMENT	Nature du jugement à l'encontre de l'entreprise
RES.N.NEGATIF	Est-ce que le résultat technique de l'entreprise en 2018 est négatif ?
TRANCHE_CA_1	Tranche du chiffre d'affaires de l'entreprise en 2018

Les différentes modalités de ces variables sont renseignées dans l'annexe 2 du mémoire.

Chapitre 3

Cadre théorique

Avant de présenter les modèles classiques et d'apprentissage statistique, nous allons introduire dans ce chapitre quelques notions fondamentales qui sont utilisées dans le mémoire.

3.1 Apprentissage supervisé et non supervisé

On parle d'apprentissage supervisé lorsque les données sont annotées d'une sortie pour entraîner le modèle, soit un label ou une classe cible. Le but de l'apprentissage supervisé est d'entraîner un algorithme capable de prédire cette cible sur des variables non annotées.

On parle d'apprentissage non supervisé lorsque les données ne sont pas annotées. L'algorithme s'applique alors à trouver des similarités dans les données d'entrée et à les regrouper dans des groupes homogènes.

Dans cette étude nous sommes dans un cas d'apprentissage supervisé où la variable à expliquer est une variable binaire :

$$DEFAUT = \begin{cases} 1 \text{ si succès (entreprise en cessation d'activité)} \\ 0 \text{ si échec (entreprise active)} \end{cases}$$

Dans cette étude la variable $DEFAUT$ est vue à la date d'extraction du répertoire Sirene, soit au 01/07/2019.

3.2 Validation croisée

Il est courant d'avoir recours à la validation croisée pour optimiser les paramètres d'un modèle (par exemple le nombre d'arbres d'une forêt aléatoire).

La validation croisée consiste à entraîner le modèle avec différents paramètres sur une base d'apprentissage, puis d'évaluer ses performances sur un échantillon indépendant appelé base de test.

L'objectif est de trouver les paramètres optimaux du modèle, c'est-à-dire ceux qui minimisent l'erreur sur l'échantillon de validation pour éviter le sur-apprentissage.

Les méthodes de validation croisée les plus utilisées sont les suivantes :

— *Holdout Method :* l'échantillon original est divisée en 2 échantillons : une base d'apprentissage (souvent de taille supérieure à 60 %) et une base de test. L'idée est de calibrer les paramètres du modèle sur la base d'apprentissage puis de calculer son erreur résultante sur la base de test.
L'erreur souvent utilisée est l'erreur moyenne quadratique pour la régression et le taux de bonnes prédictions en classification.
— *k-fold cross-validation :* on divise d'abord l'échantillon original en k sous-échantillons. On sélectionne ensuite $k-1$ échantillons qui constitueront l'ensemble d'apprentissage et l'échantillon restant servira d'ensemble de validation. On répète ainsi cette procédure k fois et on obtient k scores d'erreur. Finalement l'erreur globale estimée sera la moyenne de ces k erreurs calculées.
— *Leave One Out Cross Validation :* il s'agit du cas particulier de la *k-fold cross-validation* lorsque le nombre de sous-échantillons créés est le nombre de données à disposition, i.e. : $k = n$.

Dans cette étude nous avons le plus souvent utilisé la méthode *k-fold cross-validation* avec $k = 5$.

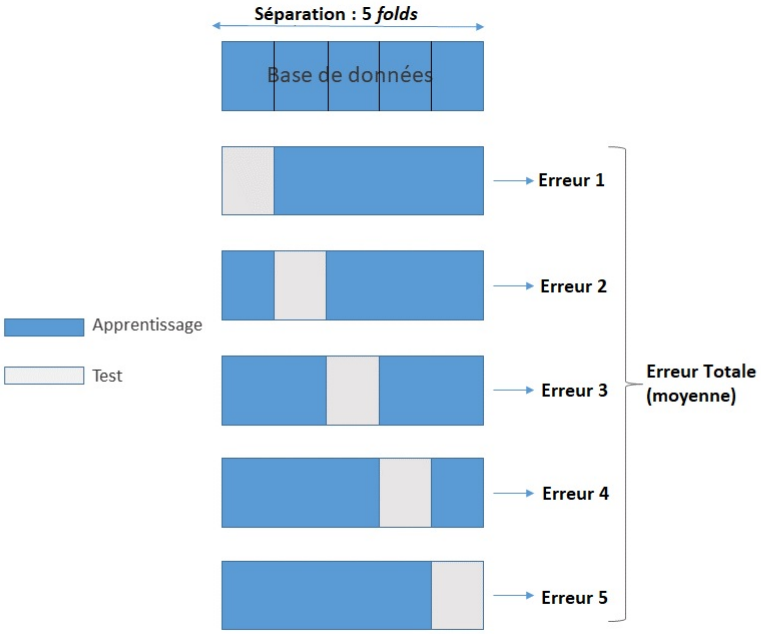

FIGURE 3.1 – *k-fold cross-validation* avec $k = 5$

3.3 Performance et sélection d'un modèle

3.3.1 Dilemme Interprétabilité / Précision

Avant de choisir notre modèle final pour modéliser la cessation d'activité nous devons considérer le dilemme *interprétabilité / performance* : allons-nous privilégier la performance ou l'interprétabilité pour choisir notre modèle ?

En effet les algorithmes ensemblistes comme la forêt aléatoire ou le XGBoost sont souvent plus performants que des modèles type GLM ou arbre CART mais ils sont

beaucoup moins interprétables (par exemple le GLM donne une équation entre la variable à expliquer et les variables explicatives).

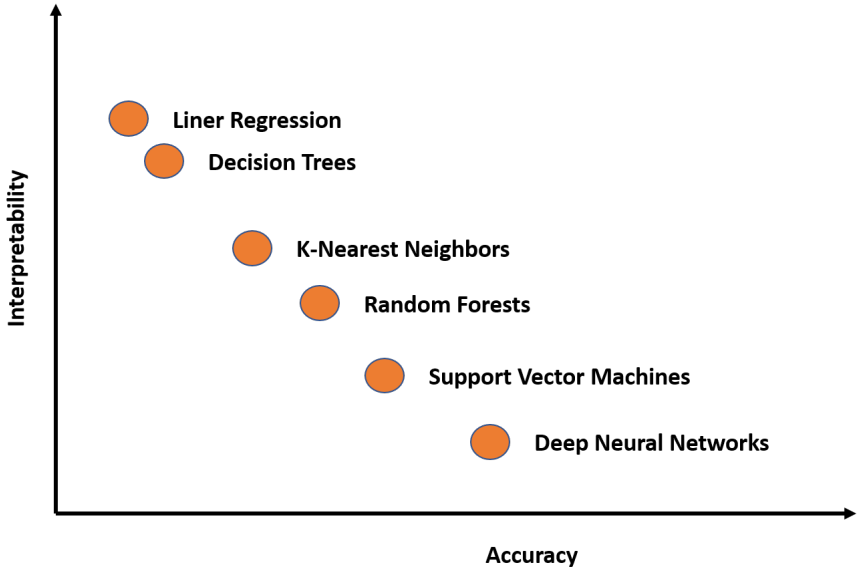

FIGURE 3.2 – Dilemme Interprétabilité / Précision (issu de l'article [10])

Dans ce mémoire nous traitons ce dilemme interprétabilité / performance de sorte à répondre la problématique métier suivante : il faut que le score de cessation d'activité soit le plus performant possible pour piloter la souscription de l'agent Generali avec le plus de précision possible. L'agent n'a pas besoin de connaître le détail de la modélisation.

C'est pourquoi le modèle de cessation d'activité sera uniquement choisi sur des critères de performance mais une partie de ce mémoire sera consacrée à l'interprétation du modèle retenu.

Vu que dans cette étude nous sommes dans un cas de classification binaire, nous allons maintenant aborder les méthodes d'évaluation associées.

3.3.2 Évaluation de la performance des modèles

Nous évaluerons nos modèles de classification sur 3 indicateurs de performance : la précision pondérée, le lift à 20% et l'Area Under the ROC Curve (AUC).

La précision pondérée

On considère la matrice de confusion suivante dans le cas d'une classification binaire.

		Observé	
		1	0
Prédit	1	Vrais Positifs (VP)	Faux Positifs (FP)
	0	Faux Négatifs (FN)	Vrais Négatifs (VN)

FIGURE 3.3 – Matrice de confusion

On définit alors la précision pondérée de la façon suivante :

$$\text{Précision pondérée} = \frac{1}{2} \times \left(\frac{VP}{VP + FP} + \frac{VN}{VN + FN}\right)$$

Si le modèle prédit aussi bien les entreprises en cessation que les entreprises encore en activité, alors la précision pondérée est équivalente à la mesure de précision classique (nombre de prédictions correctes divisé par le nombre de prédictions).

Cependant, dans un cas de données déséquilibrées où le modèle ne prédit que la classe majoritaire, la précision classique sera élevée alors que la précision pondérée sera faible.

Vu que nous avons $23,2\%$ d'entreprises en cessation dans notre base de données, nous préférons utiliser la précision pondérée pour nous assurer que le modèle prédit aussi bien les entreprises en cessation que les entreprises en activité.

La courbe lift

La courbe lift évalue la performance d'un modèle de classification sous forme visuelle.

Contrairement à la matrice de confusion qui évalue la performance d'un modèle sur toute la population, la courbe lift évalue cette performance sur une proportion de la population (on donne en général une importance au seuil 20% de la population dans la visualisation).

Dans une démarche de ciblage client, la courbe lift permet de solliciter les clients les plus réceptifs en triant la population totale par score décroissant.

Ci-dessous un exemple de lecture d'une courbe lift :

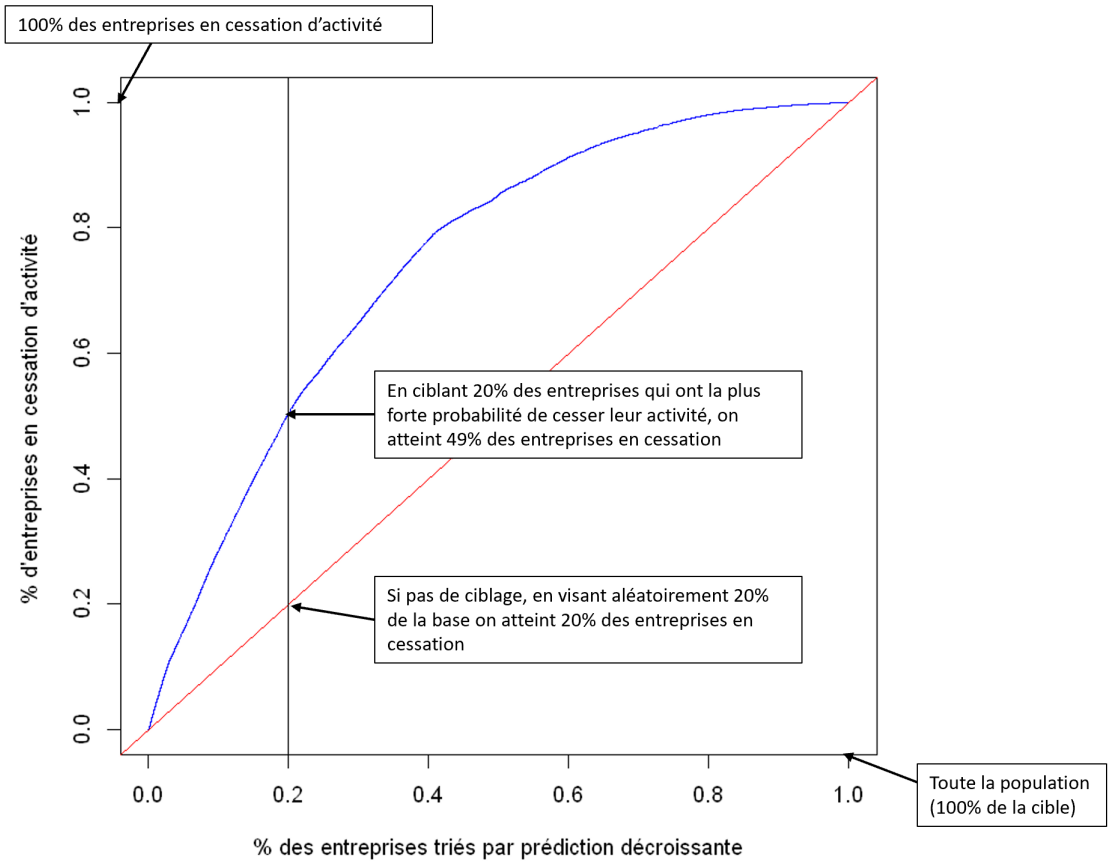

FIGURE 3.4 – Courbe lift

La courbe ROC et l'AUC

Bien que la courbe ROC soit similaire à la courbe lift, sa lecture et son interprétation sont différentes.

Graphiquement la courbe ROC donne le taux de vrais positifs (*sensibilité*) en fonction du taux de faux positifs (*spécificité*).

Dans un problème de classification binaire, notons $p_i = P(Y_i = 1)$ pour la $i^{ème}$ observation y_i.

Alors on peut choisir un seuil s à partir duquel on attribue la classe 1 à y_i si $p_i \geq s$. La courbe ROC s'obtient en faisant varier le seuil s de 1 à 0 et en calculant les taux de vrais positifs et de faux positifs pour chaque valeur de s.

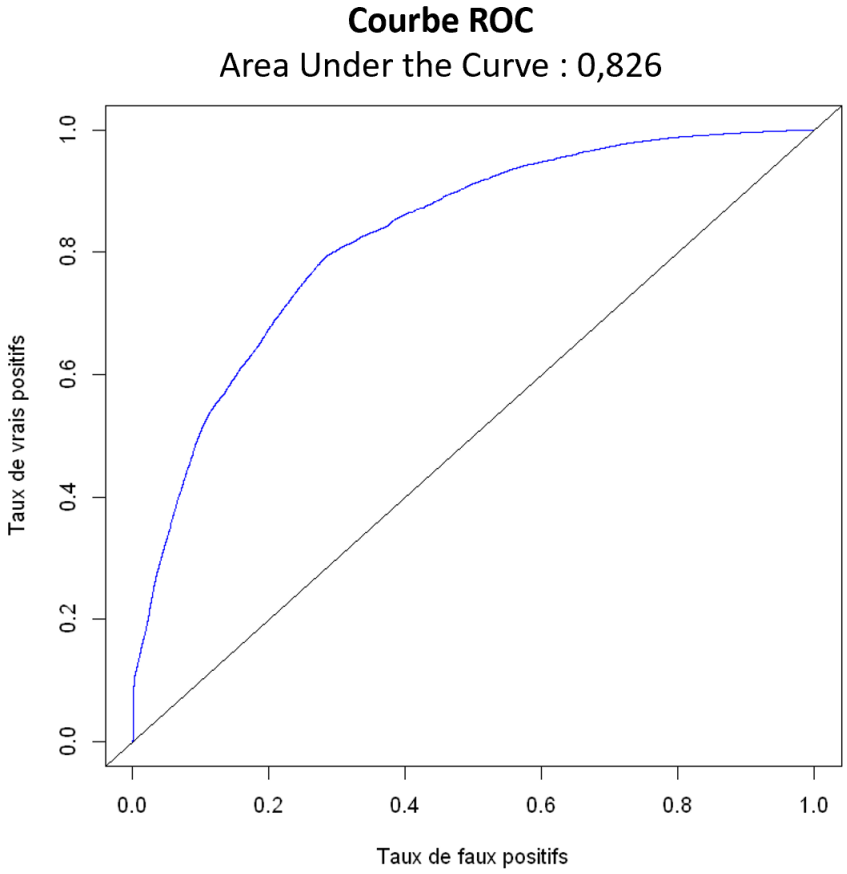

FIGURE 3.5 – Courbe ROC

On évalue alors la performance du modèle en calculant son AUC : l'aire sous la courbe ROC.

Une classification au hasard donne une AUC de 0.5 alors qu'un modèle parfait a une AUC de 1.

Chapitre 4

Modélisation

4.1 La régression logistique

Notre première méthode pour modéliser la cessation d'activité est la régression logistique. C'est une méthode couramment utilisée pour expliquer une variable dépendante qualitative Y à l'aide de variables explicatives X_i.

Dans notre étude la variable dépendante Y est binaire :

$$Y = \begin{cases} 1 \text{ si succès (entreprise en cessation d'activité)} \\ 0 \text{ si échec (entreprise active)} \end{cases}$$

Lorsque la variable dépendante Y a plus de deux modalités on parle de régression logistique polytomique.

4.1.1 Formalisation

Nous considérons l'événement « l'entreprise a cessé son activité ». Alors :

- $p = P(Y = 1)$ correspond à la probabilité que l'entreprise cesse son activité
- $1 - p = P(Y = 0)$ correspond à la probabilité que l'entreprise poursuive son activité

Soit $X = (X_1, X_2, ..., X_n)$ n variables explicatives.

Nous conditionnons alors la probabilité de cessation d'activité aux variables explicatives. On a alors :

$$\pi(x) = P(Y = 1|X = x) \text{ et } 1 - \pi(x) = P(Y = 0|X = x)$$

Désormais le modèle s'écrit :

$$g(\pi(x)) = X^t.\beta$$

avec g une fonction de lien à définir, β le vecteur des coefficients de régression. Il faut choisir la fonction g de sorte à que $g(\pi)$ soit non bornée et que π reste dans l'intervalle $[0, 1]$ après transformation. La fonction de lien utilisée dans la régression logistique est la fonction $LOGIT$ définie par $g(t) = \ln(\frac{t}{1-t})$.

Ainsi nous obtenons $g(\pi) = \ln(\frac{\pi}{1-\pi}) = \beta_0 + \sum_{j=1}^{n} \beta_j x_j$.

Par conséquent on obtient :

$$\pi(x) = \frac{exp(\beta_0 + \sum_{j=1}^{n} \beta_j x_j)}{1 + exp(\beta_0 + \sum_{j=1}^{n} \beta_j x_j)} \in [0, 1]$$

4.1.2 Odds Ratio

Dans la régression logistique l'odds ratio (ou rapport de chance) mesure l'évolution du rapport des probabilités d'apparition de l'événement $\{Y = 1\}$ par rapport à l'évenement $\{Y = 0\}$ lorsque la variable explicative X_i augmente d'une unité si elle est continue ou passe d'une modalité à une autre si elle est qualitative.

Soit $p(x) = P(Y = 1|X = x)$. Alors on écrit l'odds ratio associé à X_i la de la façon suivante :

$$OR_{X_i} = \frac{p(x_i + 1)/[1 - p(x_i + 1)]}{p(x_i)/[1 - p(x_i)]}$$

Ces odds-ratio rendent d'autant plus pertinent le découpage des variables explicatives continues en variables qualitatives pour mieux interpréter leur influence sur la variable à expliquer.

4.1.3 Estimation des paramètres

Dans les modèles de type GLM on estime les coefficients β_j avec la méthode du maximum de vraisemblance.

Soit $x(i) = (x_1, x_2, ..., x_n)$ le vecteur de réalisation de la variable X_i et $(y_1, y_2, ..., y_n)$ les observations de Y.

On cherche à maximiser la fonction suivante :

$$L_\beta = \prod_{j=1}^{n} P(Y = y_j|X_i = x_j)$$

Avec

$$P(Y = y_j|X_i = x_j) = \begin{cases} \pi(x_j) \text{ si } y_j = 1 \\ 1 - \pi(x_j) \text{ si } y_j = 0 \end{cases}$$

D'où la réécriture suivante :

$$P(Y = y_j|X_i = x_j) = \pi(x_i)^{y_i}\big(1 - \pi(x_i)\big)^{1-y_i}$$

La vraisemblance de ce modèle vaut donc $L_\beta = \prod_{i=1}^{n} \pi(x_i)^{y_i}\big(1 - \pi(x_i)\big)^{1-y_i}$.

On obtient ensuite la fonction Log-vraisemblance donnée par :

$$l_\beta = \sum_{i=1}^{n} y_i \ln(\pi(x_i)) + (1 - y_i) \ln\big(1 - \pi(x_i)\big)$$

Les paramètres sont ensuite obtenus numériquement en maximisant la fonction Log-vraisemblance avec des algorithmes de type Newton Raphson.

4.1.4 Tests statistiques de la régression logistique

Nous étudions désormais si la variable X_i est significative pour le modèle. Nous considérons alors le test statistique suivant :

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

Test du rapport de vraisemblance

Nous mesurons l'influence de la variable X_i dans le modèle par la statistique suivante :

$$G = -2\log\Big[\frac{\text{Vraisemblance sans la variable } X_i}{\text{Vraisemblance avec la variable } X_i}\Big]$$

Sous H_0, G suit asymptotiquement une loi du khi deux à un degré de liberté χ^2_1. H_0 est donc rejetée au niveau α si $G \geq \chi^2_{1-\alpha}$.

Test de Wald

Soit $\sigma(\hat{\beta})$ l'estimation de l'écart type de l'estimateur β_i. On considère la statistique suivante :

$$T = \frac{\hat{\beta}_i}{\hat{\sigma(\hat{\beta})}}$$

Sous H_0, T^2 suit une loi de khi deux à un degré de liberté χ^2_1. On rejette H_0 au niveau α si $T^2 \geq \chi^2_{1-\alpha}$.

Test du score

On calcule le score avec la formule suivante :

$$SCORE = \Big[\frac{\delta l_\beta}{\delta\beta}\Big]'_{\hat{\beta}_{H_0}} J\Big[\frac{\delta l_\beta}{\delta\beta}\Big]_{\hat{\beta}_{H_0}}$$

Avec l_β la vraisemblance, $\hat{\beta}_{H_0}$ le vecteur des paramètres estimés sous l'hypothèse H_0 et J la matrice d'information de Fisher.

4.1.5 Sélection de variables

Nous sélectionnons les variables explicatives par la méthode *Stepwise* qui conserve uniquement les variables les plus significatives sur le critère de l'AIC (*Akaike Information Criterion*) qui est définit par :

$$AIC = 2k - 2\ln(\hat{L})$$

avec k le nombre de paramètres à estimer du modèle et $\hat{L}$ sa fonction de vraisemblance maximisée.

L'AIC est donc un compromis entre le biais (qui diminue avec le nombre de paramètres) et la parcimonie (description des données avec le plus petit nombre de paramètres possible).

4.1.6 Application à nos données

Cette partie restitue les résultats de la modélisation de la cessation d'activité avec le modèle sélectionné par "méthode Stepwise".

Dans un premier temps, le test de significativité global du modèle est composé du test de Wald, du score et du rapport de vraisemblance qui sont tous significatifs au seuil 5%.

Test de l'hypothèse nulle globale : BETA=0			
Test	Khi-2	DDL	Pr > Khi-2
Rapport de vraisemblance	329 897,75	67	<,0001
Score	270 729,10	67	<,0001
Wald	189 438,13	67	<,0001

FIGURE 4.1 – Significativité du modèle

Vu que ce tableau ne nous donne pas d'indications sur la significativité individuelle de chaque variable il est nécessaire d'analyser le tableau suivant :

Analyse des effets type 3			
Variable	DLL	Khi-2 de Wald	Pr > Khi-2
Tranche d'âge de l'entreprise	4	109 540	<,0001
Tranche d'effectif de l'entreprise	4	68 636	<,0001
Nature juridique	6	21 224	<,0001
Nature du jugement à l'encontre de l'entreprise	8	15 253	<,0001
Si changement d'activité de l'entreprise	2	14 189	<,0001
Si appartenance à l'ESS	2	9 129	<,0001
Secteur d'activité de l'entreprise	11	5 343	<,0001
Si résultat négatif de l'entreprise en 2018	2	4 903	<,0001
Si l'entreprise a des employés	1	1 955	<,0001
Région du siège de l'entreprise	14	1 561	<,0001
Tranche du chiffre d'affaires de l'entreprise	3	562	<,0001
Categorie d'entreprise	2	336	<,0001
Categorie de la commune selon le zonage urbain de l'INSEE	8	51	<,0001

FIGURE 4.2 – Significativité individuelle de chaque variable

Nous observons que toutes les variables sont bien significatives au seuil 5%. De plus la variable *TRANCHE_AGE_ENT* est la plus influente sur la cessation d'activité selon le khi deux de Wald.

Observons maintenant la qualité de prédiction du modèle.

Association des probabilités prédites et des réponses observées			
% concordant	83	D de Somers	0,65
% discordant	17,5	Gamma	0,65
% lié	0	Tau-a	0,243
Nombre de paires	268 969 021 960	AUC	0,825

FIGURE 4.3 – Association des probabilités prédites et des réponses observées

Le modèle a un bon taux de bonne prédiction sur la base d'apprentissage avec une AUC à 0.825.

Afin de déterminer le seuil optimal u tel qu'une entreprise soit prédite en cessation si sa probabilité est supérieure à u pour la construction de la matrice de confusion, nous traçons la courbe de sensibilité et de spécificité contre le niveau de probabilité.

Le seuil optimal sera l'abscisse du croisement des deux courbes pour que le modèle ait la spécificité égale à la sensibilité afin d'avoir un pourcentage de bonnes prédictions de cessation égale à celui de la bonne prédiction de poursuite d'activité.

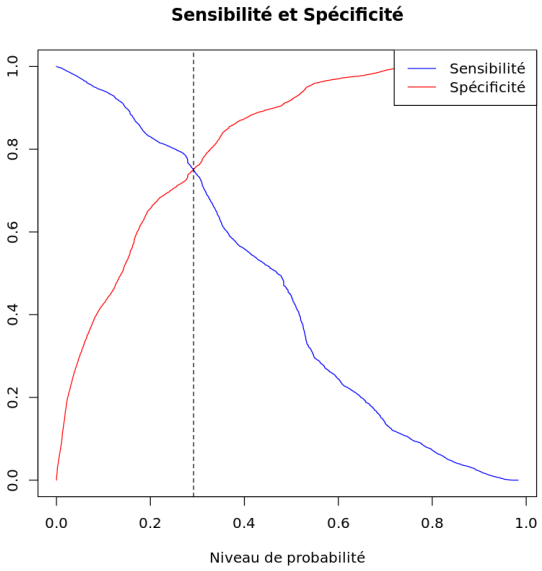

FIGURE 4.4 – Sensibilité et spécificité du modèle de régression logistique

Dans notre cas le seuil optimal u vaut $0,27$.

Pouvoir prédictif du modèle

Étudions la courbe ROC des prédictions par régression logistique.

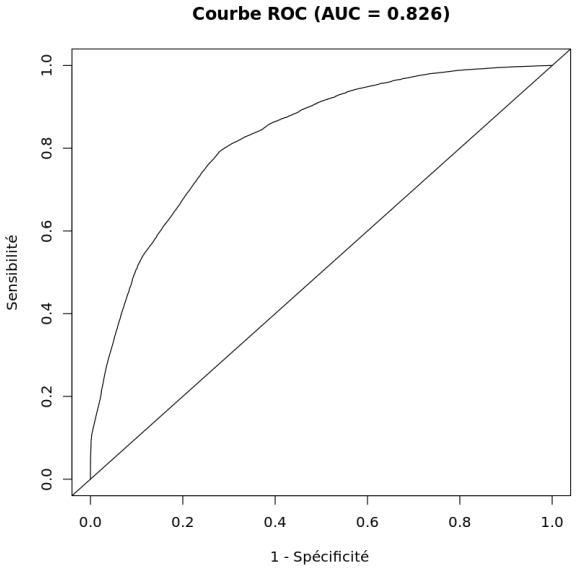

FIGURE 4.5 – Courbe ROC de la régression logistique

Le modèle est performant avec une AUC à 0.826 sur la base de test. Cette AUC est quasiment identique à l'AUC de la base d'apprentissage donc le modèle est robuste.

Ce qui nous donne la matrice de confusion suivante :

		Observé		
		0	1	Total
Prédit	0	282 271	30 868	313 139
	1	93 488	93 366	186 854
	Total	375 759	124 234	499 993

FIGURE 4.6 – Matrice de confusion de la régression logistique

On finit l'étude des prédictions par l'étude de la courbe lift sur les bases d'apprentissage et de test.

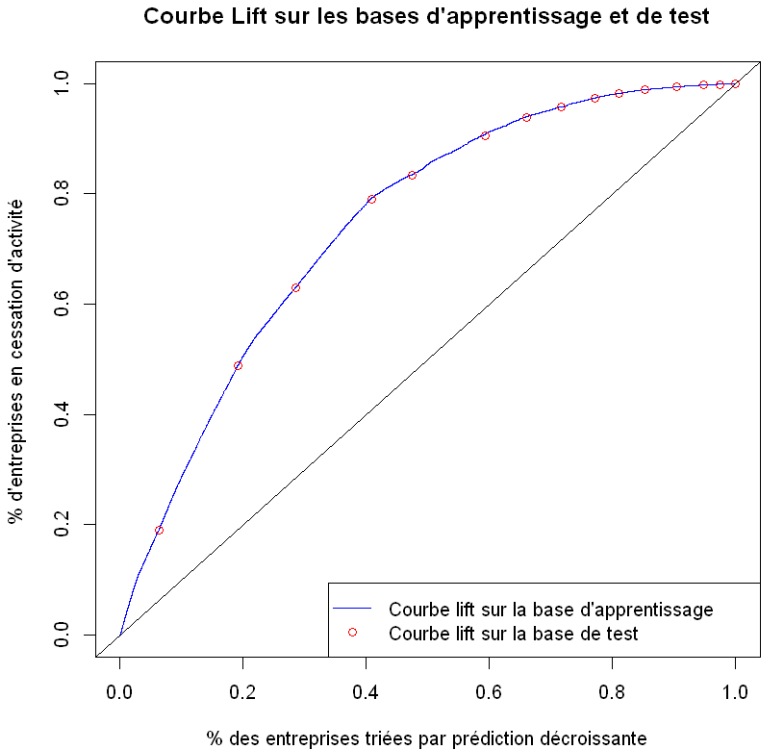

FIGURE 4.7 – Courbe lift sur la base d'apprentissage et de validation

Les deux courbes sont quasiment confondues donc le modèle est stable.

Paramètres estimés et odds-ratio

Voici les paramètres estimés de la régression logistique :

Analyse des estimations par maximum de vraisemblance						
Variables	Modalité	DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > Khi-2
Constante		1	-5,35	1,77	9,14	0,0025
Tranche d'effectif de l'entreprise	0 salarié	1	0,72	0,01	2 357,93	<,0001
	1 ou 2 salariés	1	-0,74	0,01	3 693,47	<,0001
	3 salariés ou +	1	-0,92	0,02	3 440,37	<,0001
	Effectif Inconnu	1	-0,73	0,01	6 044,05	<,0001
Tranche du chiffre d'affaires de l'entreprise	(1e+05,5,74e+10]	1	-0,10	0,01	157,34	<,0001
	(5,89e+04,8,45e+04]	1	-0,04	0,01	45,63	<,0001
	(8,45e+04,1e+05]	1	0,15	0,01	485,43	<,0001
Si appartenance à l'ESS	INCONNU	1	1,42	0,05	924,79	<,0001
	N	1	-1,63	0,05	1 069,51	<,0001
Tranche d'âge de l'entreprise	+ de 16 ans	1	-1,98	0,01	67 181,59	<,0001
	- de 3 ans	1	1,81	0,01	91 429,25	<,0001
	11 à 15 ans	1	-0,99	0,01	15 847,66	<,0001
	3 à 5 ans	1	1,07	0,01	36 781,13	<,0001
Si changement d'activité de l'entreprise	0	1	-0,50	0,04	172,71	<,0001
	1	1	0,22	0,04	32,74	<,0001
Catégorie d'entreprise	ETI	1	0,11	0,03	9,14	0,0025
	GE	1	0,33	0,04	72,87	<,0001
Si l'entreprise a des employés	N	1	0,16	0,00	1 955,20	<,0001
Nature juridique	Artisan	1	0,61	0,03	433,27	<,0001
	Artisan-commerçant	1	1,18	0,03	1 416,09	<,0001
	Autre personne morale immatriculée au RCS	1	-0,04	0,05	0,64	0,4254
	Commerçant	1	0,60	0,03	423,24	<,0001
	Groupement de droit privé	1	-1,24	0,15	68,59	<,0001
	Personne morale de droit étranger	1	-0,90	0,08	142,60	<,0001
Région du siège de l'entreprise	Auvergne-Rhône-Alpes	1	0,10	0,01	79,42	<,0001
	Bourgogne-Franche-Comté	1	0,10	0,02	44,28	<,0001
	Bretagne	1	0,08	0,01	31,21	<,0001
	Centre-Val de Loire	1	0,26	0,02	277,24	<,0001
	Corse	1	-0,31	0,03	115,40	<,0001
	Grand Est	1	0,08	0,01	39,50	<,0001
	Hauts-de-France	1	0,09	0,01	51,48	<,0001
	Inconnu	1	-0,24	0,12	4,25	0,0393
	Normandie	1	0,04	0,01	8,64	0,0033
	Nouvelle-Aquitaine	1	-0,01	0,01	0,79	0,3732
	Occitanie	1	0,06	0,01	22,28	<,0001
	Outre-Mer	1	-0,21	0,02	188,98	<,0001
	Pays de la Loire	1	0,11	0,01	57,27	<,0001
	Provence-Alpes-Côte d'Azur	1	-0,05	0,01	20,17	<,0001
Catégorie de la commune selon le zonage urbain de l'INSEE	Autre commune multipolarisée	1	0,03	0,01	3,61	0,0575
	Commune appartenant à la couronne d'un grand pôle	1	0,02	0,01	1,55	0,2129
	Commune appartenant à la couronne d'un moyen pôle	1	0,04	0,03	1,67	0,1967
	Commune appartenant à la couronne d'un petit pôle	1	0,03	0,06	0,29	0,5921
	Commune appartenant à un grand pôle (10 000 emplois ou plus)	1	0,02	0,01	3,20	0,0736
	Commune appartenant à un moyen pôle (5 000 à moins de 10 000 emplois)	1	0,04	0,02	3,99	0,0458
	Commune appartenant à un petit pôle (de 1 500 à moins de 5 000 emplois)	1	0,07	0,02	19,66	<,0001
	Commune isolée hors influence des pôles	1	0,09	0,02	29,10	<,0001
	Commune multipolarisée des grandes aires urbaines	1	0,05	0,02	2,38	0,1964
Si résultat négatif de l'entreprise en 2018	Confidentiel	1	1,08	0,02	4 387,33	<,0001
	FALSE	1	-0,91	0,02	1 416,09	<,0001
Nature du jugement à l'encontre de l'entreprise	Arrêt de la Cour d'Appel	1	0,28	1,80	0,02	0,8765
	Avis de dépôt	1	0,30	1,77	0,03	0,8651
	Extrait de jugement	1	0,58	1,77	0,11	0,7448
	INCONNU	1	1,83	1,77	1,08	0,2997
	Jugement d'ouverture	1	0,28	1,77	0,02	0,8765
	Jugement de clôture	1	0,31	1,77	0,03	0,8623
	Jugement prononçant	1	-0,01	1,77	0,00	0,9936
	Loi de 1967	1	-4,59	14,14	0,11	0,7453
Secteur d'activité de l'entreprise	ACTIVITÉS FINANCIÈRES ET D'ASSURANCE	1	0,22	0,02	171,43	<,0001
	ACTIVITÉS IMMOBILIÈRES	1	0,25	0,01	368,83	<,0001
	AGRICULTURE, SYLVICULTURE ET PÊCHE	1	-0,40	0,05	73,36	<,0001
	AUTRES ACTIVITÉS DE SERVICES	1	-0,14	0,02	81,07	<,0001
	COMMERCE ET RÉPARATION AUTOMOBILE	1	0,15	0,01	300,28	<,0001
	CONSTRUCTION	1	0,03	0,01	5,02	0,025
	ENSEIGNEMENT, SANTÉ HUMAINE, ACTION SOCIALE ET SERVICES AUX MÉNAGES	1	0,14	0,02	36,49	<,0001
	HÉBERGEMENT ET RESTAURATION	1	0,07	0,01	44,95	<,0001
	INDUSTRIE	1	0,13	0,01	139,57	<,0001
	INFORMATION ET COMMUNICATION	1	0,25	0,02	241,66	<,0001
	SOUTIEN AUX ENTREPRISES	1	0,06	0,01	58,28	<,0001

FIGURE 4.8 – Paramètres estimés de la régression logistique

Certaines modalités de variables ne sont pas significatives au seuil de 5%.

Dans un second temps nous allons analyser les odds-ratio du modèle de régression logistique. Cette étude des odds-ratio revient à l'étude de la probabilité de cessation d'activité par rapport à une modalité de référence.

Voici le tableau des odds-ratios.

Estimation des odds-ratio			
Effect	Estimation	Intervalle de confiance de Wald à 95%	
Tranche d'âge de l'entreprise 0 salarié vs Non employeur	0,39	0,37	0,40
Tranche d'âge de l'entreprise 1 ou 2 salariés vs Non employeur	0,09	0,09	0,09
Tranche d'âge de l'entreprise 3 salariés ou + vs Non employeur	0,08	0,07	0,08
Tranche d'âge de l'entreprise Effectif Inconnu vs Non employeur	0,09	0,09	0,09
Tranche du chiffre d'affaires de l'entreprise (1e+05,5,74e+10] vs [-8,8e+06,5,89e+04]	0,91	0,89	0,94
Tranche du chiffre d'affaires de l'entreprise (5,89e+04,8,45e+04] vs [-8,8e+06,5,89e+04]	0,98	0,96	0,99
Tranche du chiffre d'affaires de l'entreprise (8,45e+04,1e+05] vs [-8,8e+06,5,89e+04]	1,18	1,15	1,21
Si appartenance à l'ESS INCONNU vs O	3,32	2,54	4,33
Si appartenance à l'ESS N vs O	0,16	0,12	0,21
Tranche d'âge de l'entreprise + de 16 ans vs 6 à 10 ans	0,13	0,12	0,13
Tranche d'âge de l'entreprise - de 3 ans vs 6 à 10 ans	5,60	5,51	5,69
Tranche d'âge de l'entreprise 11 à 15 ans vs 6 à 10 ans	0,34	0,34	0,35
Tranche d'âge de l'entreprise 3 à 5 ans vs 6 à 10 ans	2,69	2,65	2,73
Si changement d'activité de l'entreprise 0 vs Inconnu	0,46	0,36	0,57
Si changement d'activité de l'entreprise 1 vs Inconnu	0,94	0,75	1,17
Catégorie d'entreprise ETI vs PME	1,72	1,58	1,88
Catégorie d'entreprise GE vs PME	2,16	1,94	2,40
Si l'entreprise a des employés N vs O	1,38	1,36	1,40
Nature juridique Artisan vs Société commerciale	2,29	2,25	2,32
Nature juridique Artisan-commerçant vs Société commerciale	4,05	3,93	4,18
Nature juridique Autre personne morale immatriculée au RCS vs Société commerciale	1,19	1,08	1,32
Nature juridique Commerçant vs Société commerciale	2,27	2,23	2,31
Nature juridique Groupement de droit privé vs Société commerciale	0,36	0,26	0,50
Nature juridique Personne morale de droit étranger vs Société commerciale	0,51	0,43	0,59
Région du siège de l'entreprise Auvergne-Rhône-Alpes vs Île-de-France	1,21	1,19	1,23
Région du siège de l'entreprise Bourgogne-Franche-Comté vs Île-de-France	1,21	1,18	1,24
Région du siège de l'entreprise Bretagne vs Île-de-France	1,19	1,16	1,22
Région du siège de l'entreprise Centre-Val de Loire vs Île-de-France	1,42	1,38	1,46
Région du siège de l'entreprise Corse vs Île-de-France	0,80	0,75	0,85
Région du siège de l'entreprise Grand Est vs Île-de-France	1,18	1,16	1,21
Région du siège de l'entreprise Hauts-de-France vs Île-de-France	1,20	1,17	1,22
Région du siège de l'entreprise Inconnu vs Île-de-France	0,86	0,67	1,10
Région du siège de l'entreprise Normandie vs Île-de-France	1,14	1,11	1,17
Région du siège de l'entreprise Nouvelle-Aquitaine vs Île-de-France	1,08	1,06	1,10
Région du siège de l'entreprise Occitanie vs Île-de-France	1,16	1,13	1,18
Région du siège de l'entreprise Outre-Mer vs Île-de-France	0,89	0,87	0,91
Région du siège de l'entreprise Pays de la Loire vs Île-de-France	1,22	1,19	1,25
Région du siège de l'entreprise Provence-Alpes-Côte d'Azur vs Île-de-France	1,04	1,02	1,06
Catégorie de la commune selon le zonage urbain de l'INSEE Autre commune multipolarisée vs Inconnu	1,43	1,13	1,81
Catégorie de la commune selon le zonage urbain de l'INSEE Commune appartenant à la couronne d'un grand pôle vs Inconnu	1,41	1,12	1,78
Catégorie de la commune selon le zonage urbain de l'INSEE Commune appartenant à la couronne d'un moyen pôle vs Inconnu	1,45	1,12	1,87
Catégorie de la commune selon le zonage urbain de l'INSEE Commune appartenant à la couronne d'un petit pôle vs Inconnu	1,43	1,06	1,94
Catégorie de la commune selon le zonage urbain de l'INSEE Commune appartenant à un grand pôle (10 000 emplois ou plus) vs Inconnu	1,42	1,12	1,79
Catégorie de la commune selon le zonage urbain de l'INSEE Commune appartenant à un moyen pôle (5 000 à moins de 10 000 emplois) vs Inconnu	1,44	1,14	1,83
Catégorie de la commune selon le zonage urbain de l'INSEE Commune appartenant à un petit pôle (de 1 500 à moins de 5 000 emplois) vs Inconnu	1,50	1,18	1,90
Catégorie de la commune selon le zonage urbain de l'INSEE Commune isolée hors influence des pôles vs Inconnu	1,51	1,20	1,92
Si résultat négatif de l'entreprise en 2018 Confidentiel vs TRUE	3,48	3,25	3,74
Si résultat négatif de l'entreprise en 2018 FALSE vs TRUE	0,47	0,43	0,52
Nature du jugement à l'encontre de l'entreprise Arrêt de la Cour d'Appel vs Rétractation sur tierce opposition	0,47	0,14	1,54
Nature du jugement à l'encontre de l'entreprise Avis de dépôt vs Rétractation sur tierce opposition	0,48	0,18	1,27
Nature du jugement à l'encontre de l'entreprise Extrait de jugement vs Rétractation sur tierce opposition	0,63	0,24	1,68
Nature du jugement à l'encontre de l'entreprise INCONNU vs Rétractation sur tierce opposition	2,22	0,84	5,89
Nature du jugement à l'encontre de l'entreprise Jugement d'ouverture vs Rétractation sur tierce opposition	0,47	0,18	1,24
Nature du jugement à l'encontre de l'entreprise Jugement de clôture vs Rétractation sur tierce opposition	0,48	0,18	1,28
Nature du jugement à l'encontre de l'entreprise Jugement prononçant vs Rétractation sur tierce opposition	0,35	0,13	0,93
Nature du jugement à l'encontre de l'entreprise Loi de 1967 vs Rétractation sur tierce opposition	0,33	0,12	0,86
Secteur d'activité de l'entreprise ACTIVITÉS FINANCIÈRES ET D'ASSURANCE vs TRANSPORTS ET ENTREPOSAGE	2,66	2,55	2,77
Secteur d'activité de l'entreprise ACTIVITÉS IMMOBILIÈRES vs TRANSPORTS ET ENTREPOSAGE	2,74	2,64	2,83
Secteur d'activité de l'entreprise AGRICULTURE, SYLVICULTURE ET PÊCHE vs TRANSPORTS ET ENTREPOSAGE	1,44	1,30	1,59
Secteur d'activité de l'entreprise AUTRES ACTIVITÉS DE SERVICES vs TRANSPORTS ET ENTREPOSAGE	1,86	1,79	1,94
Secteur d'activité de l'entreprise COMMERCE ET RÉPARATION AUTOMOBILE vs TRANSPORTS ET ENTREPOSAGE	2,49	2,42	2,57
Secteur d'activité de l'entreprise CONSTRUCTION vs TRANSPORTS ET ENTREPOSAGE	2,19	2,13	2,26
Secteur d'activité de l'entreprise ENSEIGNEMENT, SANTÉ HUMAINE, ACTION SOCIALE ET SERVICES AUX MÉNAGES vs TRANSPORTS ET ENTREPOSAGE	2,46	2,33	2,59
Secteur d'activité de l'entreprise HÉBERGEMENT ET RESTAURATION vs TRANSPORTS ET ENTREPOSAGE	2,29	2,22	2,36
Secteur d'activité de l'entreprise INDUSTRIE vs TRANSPORTS ET ENTREPOSAGE	2,44	2,37	2,51
Secteur d'activité de l'entreprise INFORMATION ET COMMUNICATION vs TRANSPORTS ET ENTREPOSAGE	2,73	2,63	2,85
Secteur d'activité de l'entreprise SOUTIEN AUX ENTREPRISES vs TRANSPORTS ET ENTREPOSAGE	2,28	2,22	2,34

FIGURE 4.9 – Odds-ratios de la régression logistique

Quelques interprétations :

Une entreprise âgée de - de 3 ans accuse 5,6 fois plus de chance de cesser son activité qu'une entreprise âgée de 6 à 10 ans toutes choses égales par ailleurs. Un artisan-commerçant a 4 fois plus de chances de cesser son activité qu'une société commerciale toutes choses égales par ailleurs.

4.2 L'arbre CART

Dans cette section nous allons modéliser la cessation d'activité à l'aide d'un arbre Classification And Regression Trees (CART).

Il s'agit d'une méthode statistique, introduite par Breiman et al. (1984)[3] qui construit des prédicteurs par arbre aussi bien en régression qu'en classification.

Le principe de CART est de faire une partition récursive de l'espace d'entrée à l'aide d'une règle de décision binaire, puis de déterminer une sous-partition optimale pour

la prédiction.

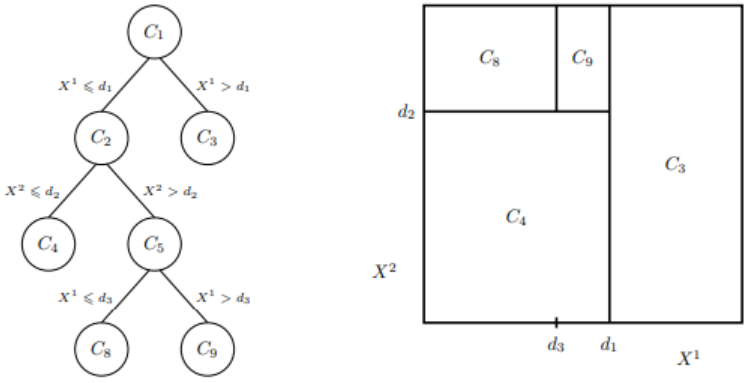

FIGURE 4.10 – A gauche : un arbre de classification qui permet de prédire la classe correspondante selon un x donné. A droite : la partition associée dans l'espace des variables explicatives (issu de l'article [8])

La construction d'un arbre CART se fait en deux étapes :

- construction d'un arbre maximal qui permet de définir la famille de modèles à l'intérieur de laquelle on cherchera à sélectionner le meilleur
- l'élagage qui construit une suite de sous-arbres optimaux élagués de l'arbre maximal

Le principal avantage des arbres de décision est qu'ils sont facilement interprétables et explicables à un utilisateur non expert.

4.2.1 Construction de l'arbre maximal

La méthode CART construit un arbre binaire dont les noeuds sont des sous-échantillons des données d'apprentissage.

- le noeud racine contient toutes les données d'apprentissage
- à chaque étape le noeud est divisé pour construire deux nouveaux noeuds les plus **homogènes** possible au sens de la variable à expliquer
- l'arbre maximal est obtenu lorsqu'aucun noeud ne peut plus être divisé. Un noeud terminal (qui ne peut plus être divisé) est appelée une **feuille**
- chaque feuille est alors associée à l'une des classes de la variable à expliquer

Mesures de qualité d'une division

Soit $X = (X_1, ..., X_p)$ les variables explicatives quantitatives ou qualitatives. Soit y la variable à expliquer à K modalités définissant les K classes à prédire.

Le but est de diviser un noeud t en deux sous-noeuds t_G (noeud fils gauche) et t_D (noeud fils droit) qui soient le plus homogène par rapport à la variable à expliquer y.

On calcule l'hétérogénéité d'un noeud à partir d'une fonction ϕ appelée **fonction d'impureté** qui doit être :

- positive ou nulle
- nulle si toutes les observations du noeud appartiennent à la même classe que y : on parle de noeud pur

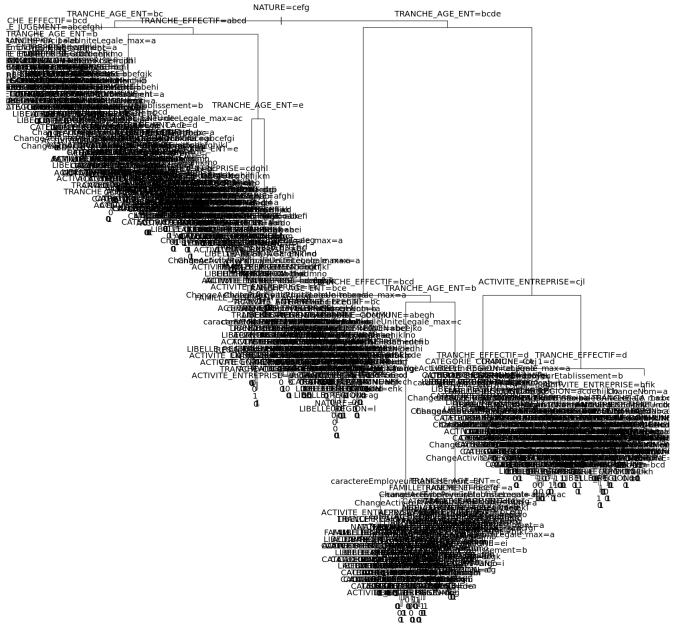

FIGURE 4.11 – Arbre maximal

— maximale lorsque les classes de y sont équiprobables dans le noeud : on parle de noeud impur

La fonction d'impureté ϕ est définie sur l'ensemble des K-uplets $(p_1, ..., p_k)$ avec $p_k \geq 0$ pour $k \in 1, ..., K$ et $\sum_{k=1}^{K} p_k = 1$ et doit respecter les propriétés suivantes :

— ϕ admet un unique maximum en $(\frac{1}{K}, ..., \frac{1}{K})$
— ϕ est minimum aux points $(1, 0, ..., 0)$, $(0, 1, ..., 0)$, ...
— ϕ est une fonction symétrique de $p_1, ..., p_k$

On définit alors l'impureté $i(t)$ d'un noeud t par :

$$i(t) = \phi(p_{t,1}, ..., p_{t,K})$$

avec $p_{t,k} = \frac{n_{t,k}}{n_t}$ la proportion de la classe k dans le noeud t. Cette proportion donne la probabilité de la classe k dans le noeud t.

Pour calculer l'impureté d'un noeud t dans un arbre CART on utilise principalement les deux mesures suivantes :

• L'indice de Gini :

$$i(t) = \sum_{k=1}^{K} p_{t,k}(1 - p_{t,k}) = 1 - \sum_{k=1}^{K} p_{t,k}^2$$

• L'entropie

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

La qualité $\delta(t_G, t_D)$ d'une division (t_G, t_D) d'un noeud t est la réduction de l'impureté obtenue par cette division :

$$\delta(t_G, t_D) = i(t) - p_G i(t_G) - p_D i(t_D)$$

où p_G et p_D sont respectivement les proportions d'observations de t partant dans le noeud gauche t_G et dans le noeud droit t_D. C'est pourquoi une bonne division occasionnera une forte diminution de l'impureté.

Division d'un noeud

L'algorithme consiste donc à choisir la division (t_G, t_D) qui maximise $\delta(t_G, t_D)$ parmi toutes les divisions possibles : on cherche à diminuer le plus possible l'impureté.

Chaque division (t_G, t_D) d'un noeud t est effectuée par des questions binaires. Une question binaire est définie à partir d'une variable explicative X_i de la manière suivante :

• Si $X_i \in \mathbb{R}$ est quantitative, la question binaire sera du type

$$X_i \leq C?$$

Il existe alors une infinité de valeurs de découpage C avec au maximum n_t divisions différentes.

• Si $X_i \in 1, ..., M$ est qualitative, la question binaire sera du type

$$X_i \subset B?$$

où $B \subset 1, ..., M$. Il existe $2^{M-1} - 1$ questions binaires donc au maximum $2^{M-1} - 1$ divisions différentes.

Le critère d'arrêt de l'algorithme consiste à ne pas découper un noeud pur ou un noeud contenant trop peu de données.

Enfin, en classification à chaque nœud terminal t de l'arbre est associé le label de la classe majoritaire des observations présentes dans le noeud t.

4.2.2 Elagage

La deuxième étape de l'algorithme CART s'appelle l'élagage et consiste à chercher le meilleur sous-arbre élagué de l'arbre maximal afin d'éviter le sur-apprentissage et obtenir un modèle plus parcimonieux.

Cette étape consiste à :

— construire une suite de sous-arbres emboîtés
— choisir le sous-arbre optimal au sens d'un critère mesurant un compromis entre la taille de l'arbre et son coût de mauvais classement

On cherche donc le coefficient α qui minimise la fonction de coût-complexité $C_\alpha(T)$ définie par :

$$C_\alpha(T) = R(T) + \alpha|T|$$

avec $|T|$ le nombre de feuilles de l'arbre, α un réel positif pénalisant la complexité de l'arbre et $R(T)$ l'erreur de mauvais classement.

Pour construire la séquence de sous-arbres emboités, il suffit de trier par ordre croissant les noeuds de l'arbre maximal T_{max} en fonction de leur paramètre de complexité, puis de supprimer successivement les divisions associées à ces noeuds.

Voici les résultats de l'élagage de notre arbre maximal :

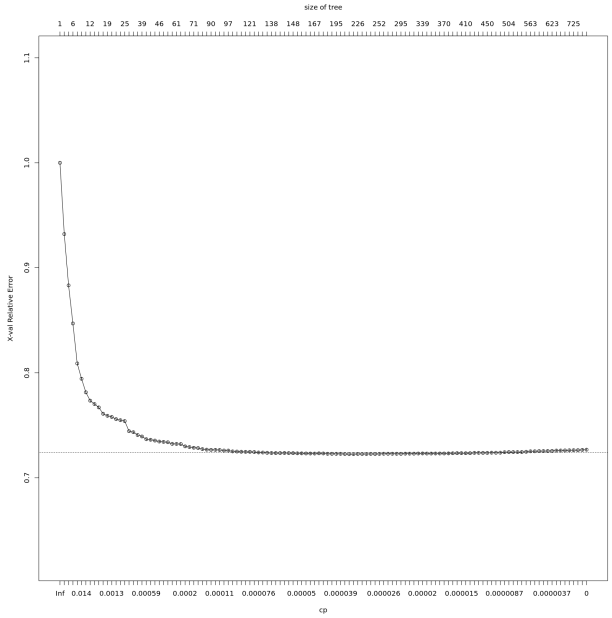

FIGURE 4.12 – Erreur par validation croisée en fonction du nombre de feuilles

L'erreur minimale est atteinte lorsque le paramètre *cp* vaut 0.0000335. On obtient alors l'arbre élagué suivant :

FIGURE 4.13 – Arbre retenu après élagage

On est alors passé d'un arbre maximal à 756 feuilles à un arbre élagué à 218 feuilles.

Pour conclure les arbres possèdent les avantages suivants :

- modèle non paramétrique : pas d'hypothèses sur la loi de y
- cadre unique pour la régression ou la classification
- modèles facilement interprétables
- gestion des variables explicatives quantitatives et qualitatives

Cependant leur principal inconvénient est le manque de stabilité : un léger changement dans la base d'apprentissage peut entraîner un fort changement dans la construction de l'arbre et donc dans les prédictions.

4.3 Prolongement des arbres CART vers des méthodes ensemblistes

Comme nous l'avons vu dans la section précédente, les arbres CART sont généralement très instables.

L'idée principale des méthodes ensemblistes est d'augmenter la fiabilité et la stabilité des résultats en agrégeant un grand nombre d'arbres et en réunissant leurs résultats. En effet on met en commun plusieurs arbres pour obtenir de meilleurs performances.

Il existe trois façons d'agréger les arbres CART pour améliorer leur performance :

- le bagging
- les forêts aléatoires
- le boosting

Pour les modèles forêt aléatoire et XGBoost nous avons optimisé les hyperparamètres par validation croisée avec le package *R caret*.

4.3.1 Le Bagging

Le *bagging* est la contraction de *bootstrap aggregating*.

L'algorithme consiste à construire plusieurs arbres par bootstrap, c'est à dire que B échantillons tirés avec remise sur la base d'apprentissage sont constitués, à partir desquels B arbres maximaux sont construits. Puis la variable à expliquer est prédite en agrégeant les B estimations obtenues à partir des B arbres construits.

Si la variable à expliquer est qualitative, un vote à la majorité est effectué parmi les estimations des B arbres, si la variable à expliquer est quantitative la moyenne des B prédictions est utlisée.

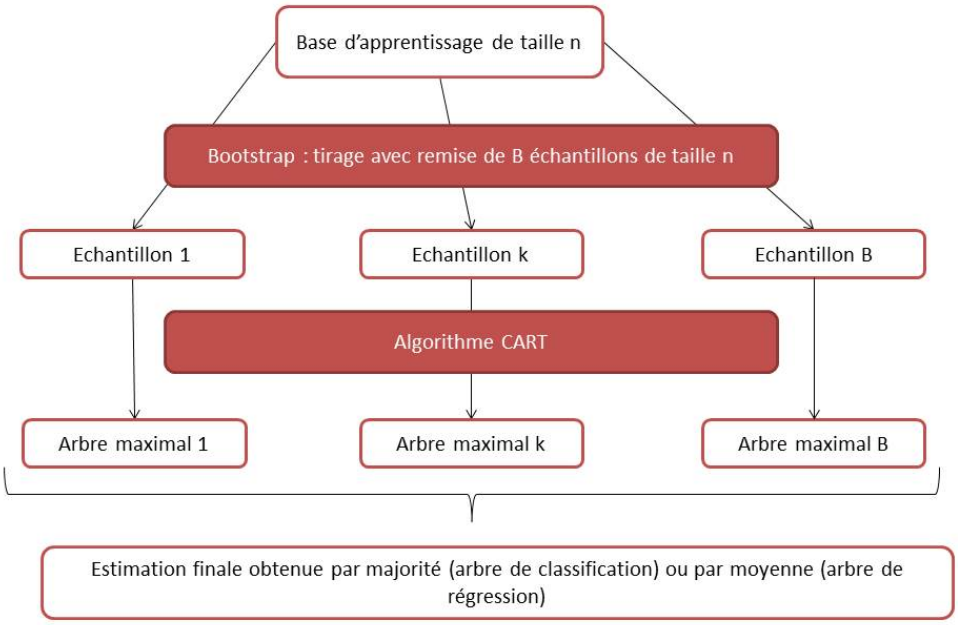

FIGURE 4.14 – Algorithme de bagging

4.3.2 Forêts aléatoires

Cette méthode a été introduite par L. Breiman en 2001[2].

Elle est similaire au *bagging* dans la manière d'agréger des modèles mais la différence se fait dans la création des arbres qui vont être agrégés. En effet dans le *bagging* à chaque division de noeud on choisit la meilleure division parmi toutes les divisions, alors que dans une forêt aléatoire on tire aléatoirement un certain nombre m (ou *mtry*) de variables explicatives, et on considère les divisions possibles basées sur ce sous-ensemble.

Approximativement m vaut la racine carrée du nombre de variables explicatives pour un arbre de classification et un tiers du nombre de variables explicatives pour un arbre de régression.

L'objectif des forêts aléatoires est de réduire la corrélation des arbres présents dans le *bagging* en les diversifiant.

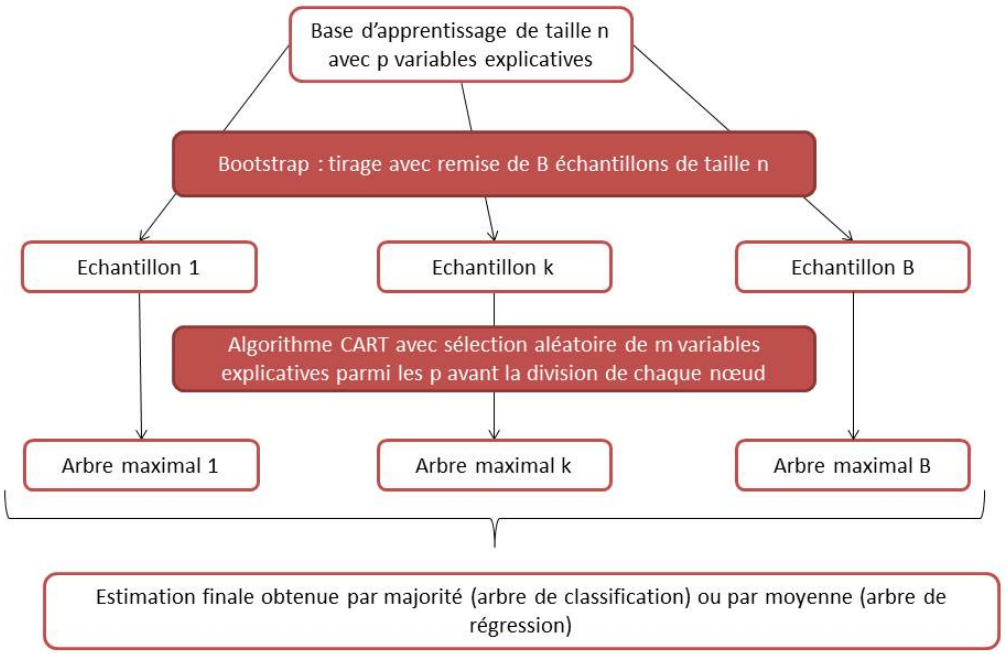

FIGURE 4.15 – Algorithme de forêt aléatoire

Erreur *Out-of-bag* (OOB)

Pour évaluer le score de performance d'une forêt aléatoire nous pouvons éviter le découpage entre base d'apprentissage et base de validation et nous servir de l'échantillon OOB car les arbres construits n'utilisent pas toutes les observations de l'échantillon d'apprentissage. En effet pour un arbre donné certains individus sont tirés plusieurs fois, tandis que d'autres ne sont pas inclus dans l'échantillon bootstrap (environ $36,8\%$). Nous pouvons alors appliquer cet arbre sur ces individus pour obtenir une prédiction.

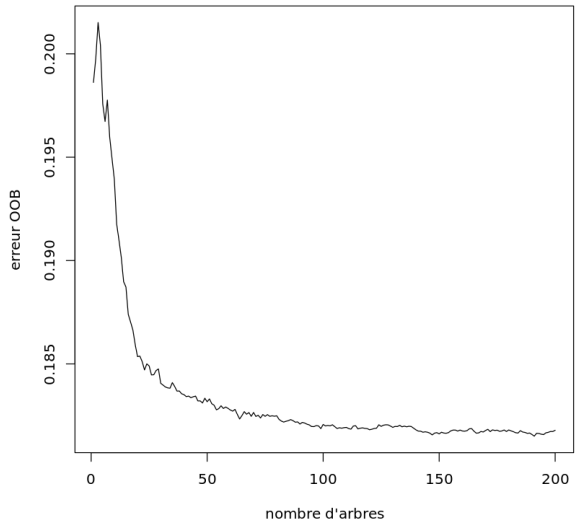

FIGURE 4.16 – Erreur du modèle en fonction du nombre d'arbres

Importance des variables par forêt aléatoire

Vu qu'une forêt aléatoire agrège des modèles, il n'y a pas d'interprétation directe possible. Cependant ce modèle permet de calculer l'importance de chaque variable en calculant son critère d'impureté de Gini.

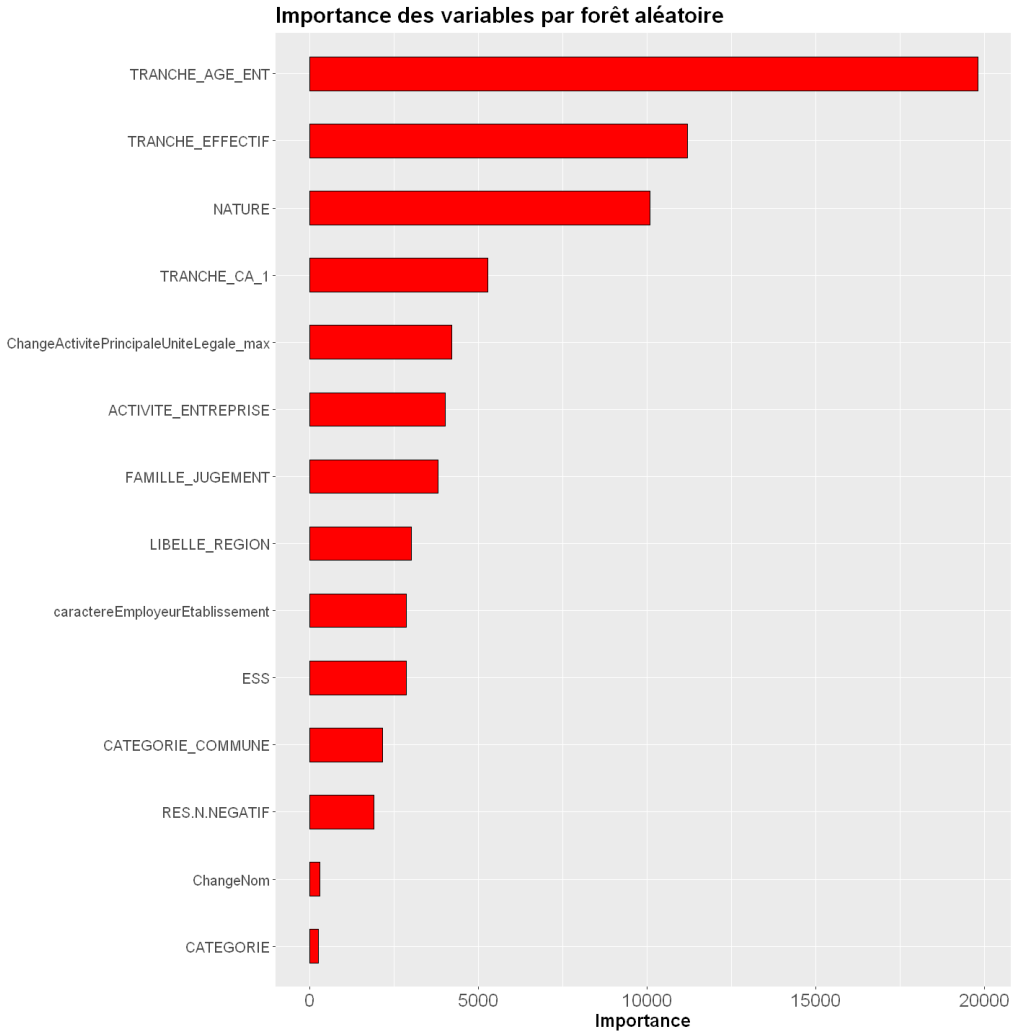

FIGURE 4.17 – Importance des variables par forêt aléatoire

Les trois variables les plus importantes pour modéliser la cessation d'activité par forêt aléatoires sont : la tranche d'âge de l'entreprise, sa tranche d'effectif et sa nature juridique. C'est en cohérence avec notre étude statistique sur les liaisons entre les variables explicatives et la variable à expliquer.

4.3.3 Boosting

Contrairement aux méthodes précédentes qui construisent des arbres indépendants les uns des autres, le *Boosting* consiste à construire les arbres les uns après les autres de façon à que chaque arbre corrige les défauts du précédent.

En effet les observations mal prédites par le $i^{ème}$ arbre vont être surpondérées dans l'arbre suivant afin de leur donner plus d'importance : l'algorithme s'améliore au fur et à mesure. Le *Boosting* concentre ainsi ses efforts sur les observations les plus difficiles à ajuster tandis que l'agrégation de l'ensemble des modèles permet d'éviter le sur-apprentissage.

Enfin il existe plusieurs algorithmes de *Boosting*. Ils diffèrent selon leur manière de pondérer pour renforcer l'apprentissage des données mal prédites, leur fonction de perte ou leur façon d'agréger les modèles successifs.

• Principe du Gradient Boosting

L'article de Friedman en 2001[7] est la référence du *gradient boosting*.

Soit $F(x)$ la fonction permettant de lier les variables explicatives $x = (x_1, ..., x_n)$ à la variable à expliquer y. C'est la fonction qui minimise la moyenne d'une fonction de perte notée L sachant l'échantillon.

$$\hat{F} = \underset{F}{\operatorname{argmin}}\, \mathbb{E}_{x,y}[L(y, F(x))]$$

L'objectif du *gradient boosting* est de trouver la meilleure estimation $\hat{F}$ de la fonction F grâce à une somme pondérée de classifieurs faibles h_i (généralement des arbres CART).

$$\hat{F} = \sum_{i=1}^{M} \gamma_i h_i(X) + cst$$

Puis le *gradient boosting* essaye de trouver l'approximation $\hat{F(X)}$ qui minimise la moyenne de la fonction de coût sur l'échantillon d'apprentissage. Il commence d'abord par prendre la fonction constante F_0 qui minimise la moyenne des erreurs, puis à chaque itération $m \in 1, ..., M$ il complexifie la fonction en utilisant celle de l'itération précédente.

En formalisant :

$$F_0(X) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$$

$$F_m(X) = F_{m-1}(X) + \underset{h_m}{\operatorname{argmin}}[\sum_{i=1}^{n} L(y_i, F_{m-1}(x_i + h_m(x_i))]$$

• Une variante du Gradient Boosting : XGBoost

Dans ce mémoire nous allons utiliser l'algorithme XGBoost. Il s'agit d'une implémentation *open source* optimisée et parallélisée du Gradient Boosting présentée en 2015 par T. Chen et C. Guestrin[5].

4.4 Sélection du modèle de cessation d'activité

Nous mesurons les performances des modèles sur la base de test, indépendante de la base d'apprentissage.

4.4.1 Première comparaison des performances des modèles

Étude des courbes ROC et des AUC

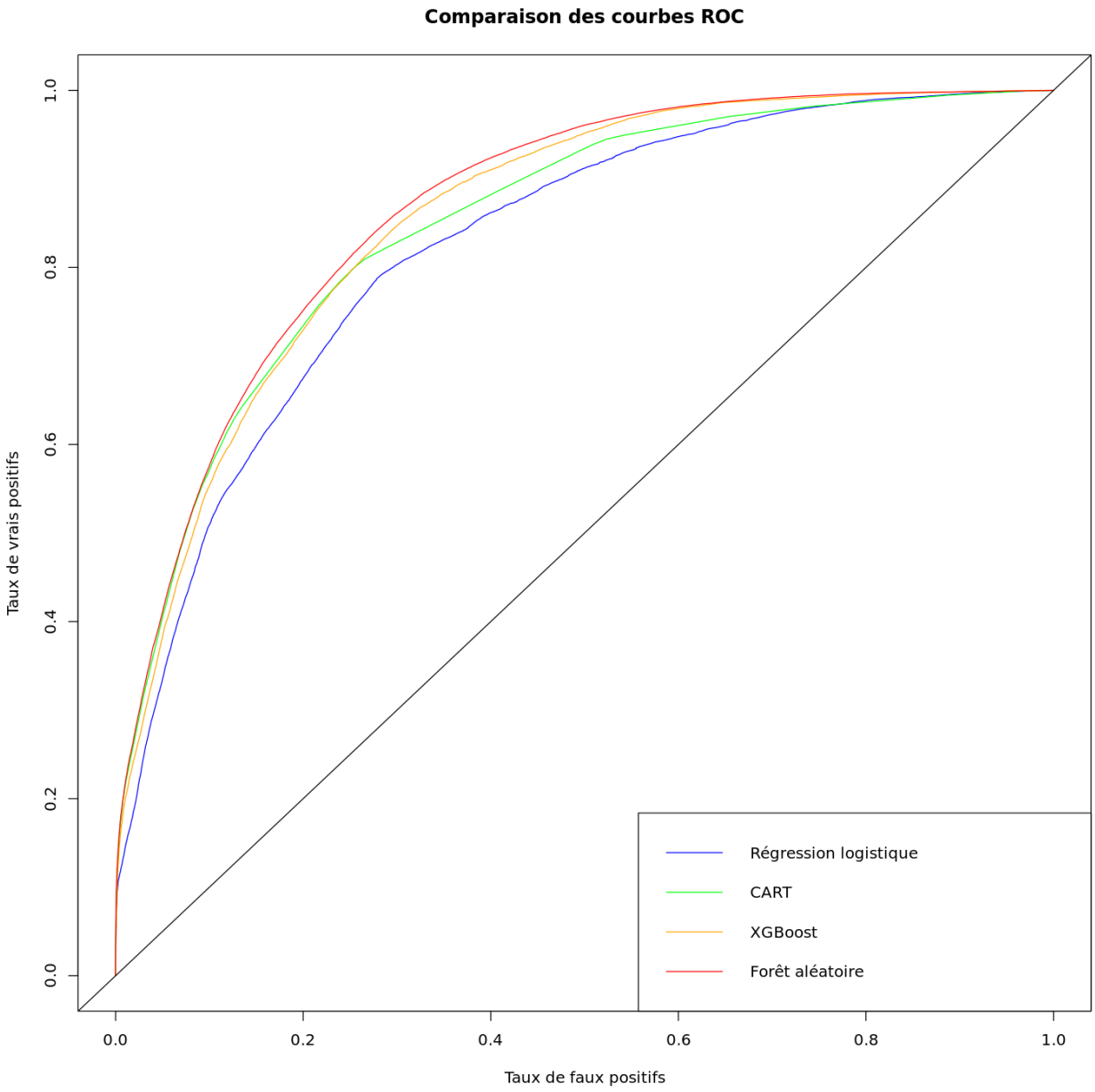

FIGURE 4.18 – Courbes ROC

Avec une courbe ROC au dessus des autres et une AUC de 0.868 le modèle forêt aléatoire est le plus performant et l'interprétation associée est qu'il y a 86, 8% de chances que le modèle distingue bien une entreprise qui a cessé son activité d'une entreprise toujours en activité. Ce modèle est donc performant.

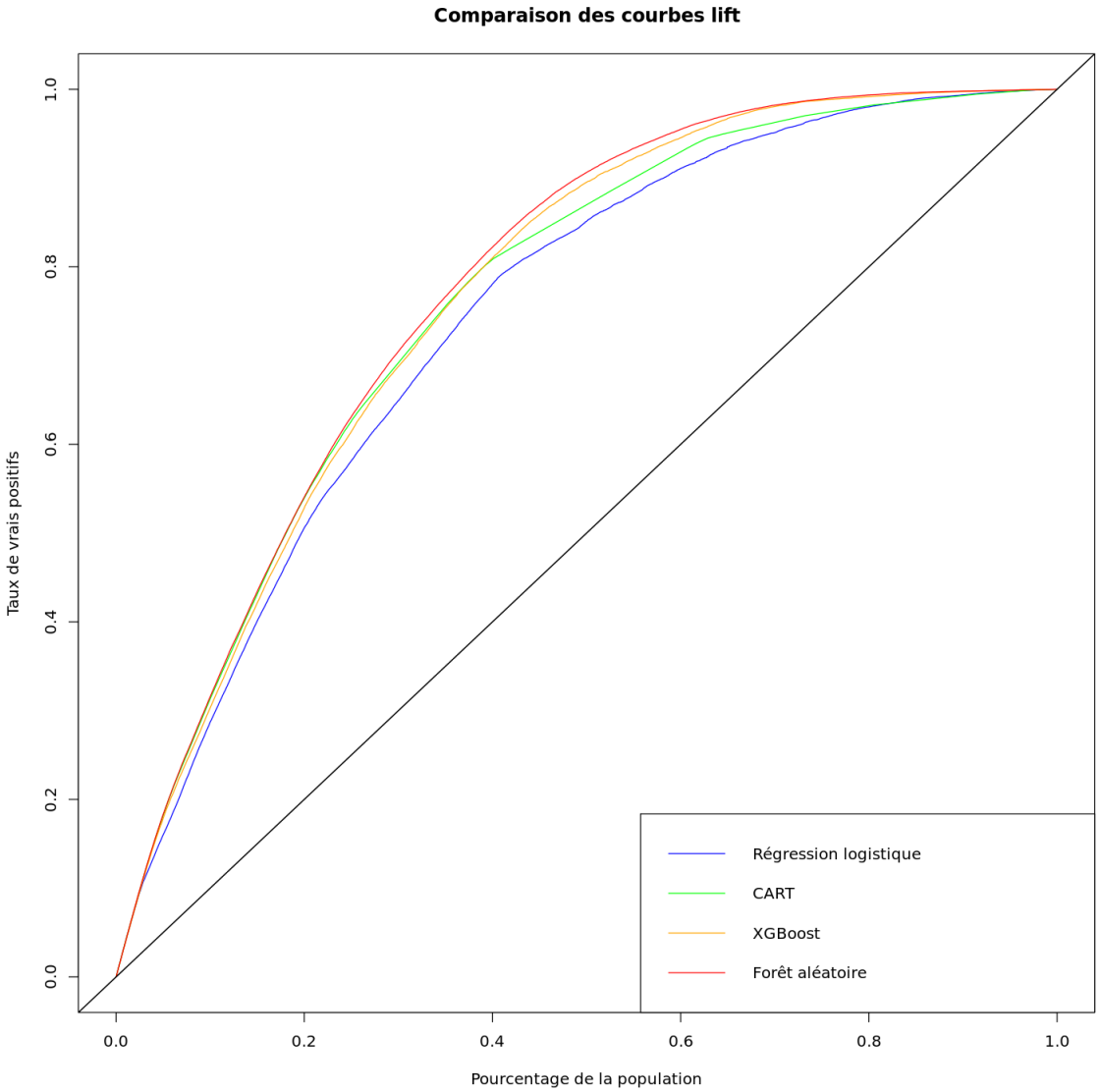

FIGURE 4.19 – Courbes lift

Étude des courbes lift

La forêt aléatoire est encore la plus performante : sa courbe lift est toujours au dessus des autres. Avec un lift à 20% à 0.531, cela signifie que $53,1\%$ des entreprises qui ont cessé leur activité sont parmi les 20% d'entreprises avec le plus haut score de cessation d'activité.

Étude de la précision pondérée

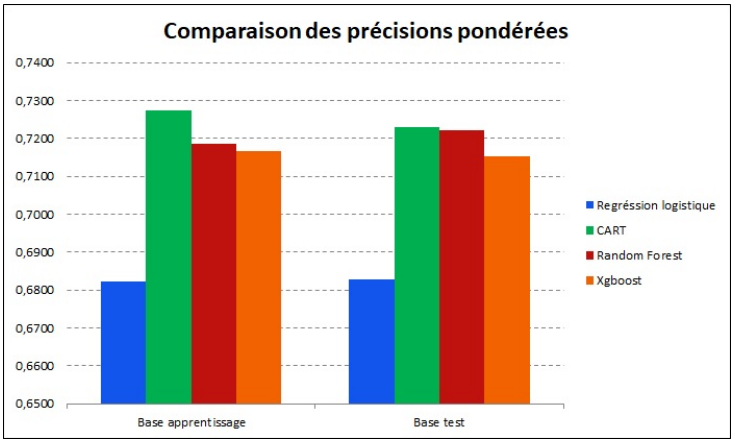

FIGURE 4.20 – Précisions pondérées des modèles de cessation

Le graphique nous montre que le GLM fait preuve d'une précision pondérée plus faible que les autres modèles et que l'arbre CART est légèrement meilleur sur la base de test.

Nous reviendrons pour commenter ces résultats dans une partie ultérieure.

En effet avant de sélectionner le modèle qui va modéliser la cessation d'activité, nous allons vérifier que les modèles présentés sont stables vis à vis des bases de test et d'apprentissage.

4.4.2 Stabilité des modèles vis à vis des bases d'apprentissage et de test

Dans les parties précédentes les modèles ont été ajustés à partir d'une base d'apprentissage obtenue en échantillonnant 70% de la base de données initiale.

Les paramètres optimaux des modèles présentés ont ensuite été calibrés sur cette base d'apprentissage par validation croisée.

Quant aux critères de performance, ils sont calculés sur la base de test, soit les 30% restant de la base de données initiale.

Cependant, bien que cette séparation entre la base d'apprentissage et la base de test soit aléatoire, il n'est pas garantit que sur un nouvel échantillon les modèles utilisés et leurs paramètres choisis soient aussi performants.

C'est pourquoi nous réalisons la même étude que précédemment en choisissant, toujours de manière aléatoire, une nouvelle base d'apprentissage et une nouvelle base de test, afin de vérifier la stabilité des résultats.

Pour ce faire, nous prenons quatre nouveaux échantillonnages de la base initiale et nous conservons les mêmes critères de performance présentés dans la partie précédente.

Nous obtenons les résultats suivants pour les bases de test des nouveaux échantillonnages :

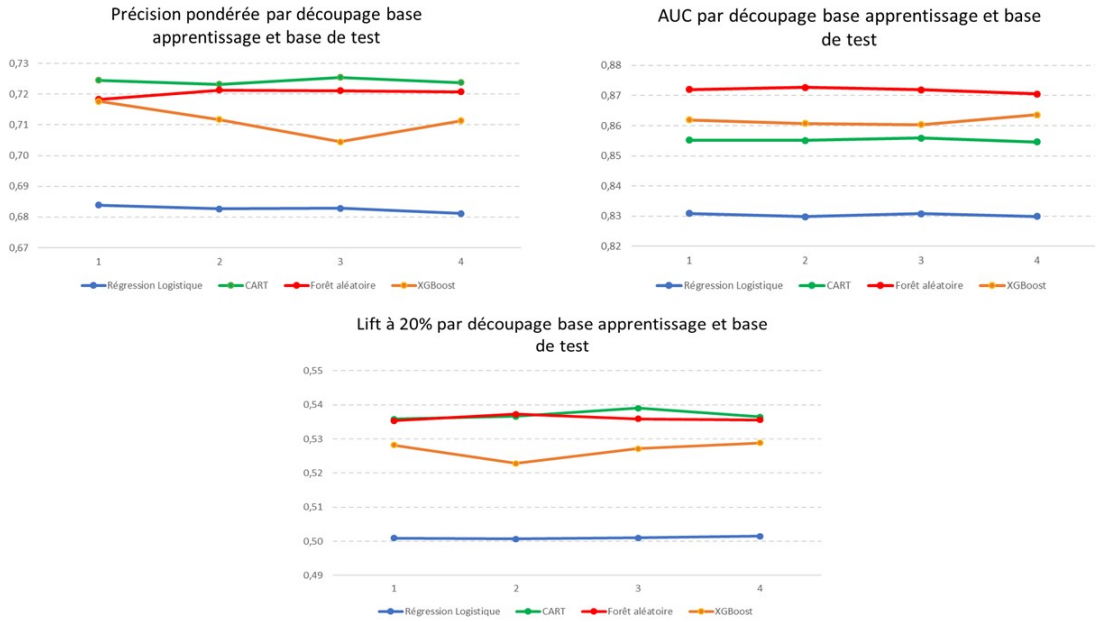

FIGURE 4.21 – Critères de performance par échantillonnage des bases d'apprentissage et de test

Au vu de ces graphiques, on remarque une certaine stabilité des critères de performance selon l'échantillonnage des bases d'apprentissage et de test.

Maintenant que nous sommes assurés que les modèles sont stables, nous nous intéressons aux pertes relatives par modèle et par critère de performance qui vont nous aider à sélectionner notre modèle de cessation d'activité.

Mais avant cela, il est nécessaire de rappeler de quelle manière nous allons choisir le modèle pour ce mémoire : nous cherchons à modéliser la cessation d'activité avec le plus de précision possible pour que les agents généraux de Generali aient à leur disposition l'indicateur le plus fiable possible.

C'est pourquoi afin d'être en adéquation avec l'utilisation métier de cet indicateur, nous allons comparer les modèles entre eux et nous sélectionnerons le meilleur au regard des critères de performance retenus.

Pour cela pour chaque modèle et critère de performance, nous considérons la moyenne sur les quatre échantillonnages, puis nous calculons les pertes relatives par rapport au modèle qui a le meilleur critère de performance moyen.

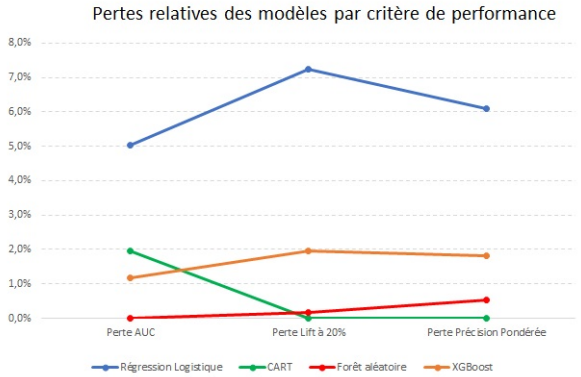

FIGURE 4.22 – Perte des modèles par critère de performance

Tout d'abord, au vu du graphique il semble naturel d'exclure la régression logistique qui accuse des pertes relatives de plus de 5% par rapport au meilleur modèle sur chaque critère de performance.

Ensuite le XGBoost accuse également des pertes entre 1% et 2% sur chaque critère de performance par rapport au meilleur modèle : nous l'excluons des modèles candidats pour modéliser la cessation d'activité.

Il ne reste que la forêt aléatoire et l'arbre CART comme modèles candidats.

Bien que ces deux modèles aient un lift à 20% et une précision pondérée quasi identiques, l'arbre CART accuse une perte relative de presque 2% en AUC : cela signifie que l'arbre CART distingue moins bien si une entreprise a cessé son activité ou non par rapport à la forêt aléatoire.

Vu que notre choix de modèle de cessation d'activité se porte sur le modèle le plus performant nous sélectionnons la forêt aléatoire.

Toutefois, même si nous avons choisi un modèle de cessation d'activité satisfaisant, nous allons nous poser la question de l'interprétabilité de ce modèle ainsi que quelques limites de cette modélisation.

4.5 Interprétation des modèles complexes : un enjeu incontournable

Dans la partie précédente nous avons sélectionné la forêt aléatoire pour modéliser la cessation d'activité en nous appuyant uniquement sur ses performances.

Or bien que ce modèle nous ait apporté un apport non négligeable en terme de performance, ce gain d'efficacité a un coût : nous perdons en lisibilité et en transparence.

En effet la forêt aléatoire fait partie de ces algorithmes considérés comme des « boîtes noires », comme également le XGBoost ou les réseaux de neurones, vu qu'ils sont difficilement interprétables contrairement à des modèles de type GLM par exemple.

Néanmoins comprendre ces modèles complexes va devenir un enjeu incontournable.

Tout d'abord pour des raisons réglementaires : l'Autorité de Contrôle Prudentiel et de Résolution (ACPR) pour l'assurance et plus généralement le Règlement Général sur la Protection des Données (RGPD) exigent des justifications sur les décisions prises par les algorithmes et l'assureur ne peut pas transférer sa responsabilité sur la machine.

Ensuite il est nécessaire d'être capable d'interpréter ces modèles pour pouvoir expliquer leurs résultats et d'apporter de la confiance dans leur utilisation.

Ainsi, c'est dans cette volonté de mieux comprendre ces algorithmes complexes que des articles sur le Machine Learning Interpretable (IML) sont publiés et que des méthodes pour interpréter ces modèles apparaissent.

Les deux méthodes d'interprétabilité les plus utilisées aujourd'hui sont les méthodes LIME et SHAP.

Dans ce mémoire nous allons uniquement utiliser la méthode LIME pour interpréter notre modèle de cessation d'activité tandis que la méthode SHAP sera en annexe pour permettre au lecteur de la consulter.

4.5.1 Méthode LIME

Tout d'abord la méthode LIME est une méthode locale d'interprétation de modèles de *machine learning* à partir de modèles de substitution : cela signifie que son objectif est d'approcher le comportement de l'algorithme complexe par un modèle simplifié comme un arbre de décision peu profond ou un modèle linéaire parcimonieux.

L'approche locale consiste à réaliser la même opération mais au niveau d'une seule observation : le comportement de l'algorithme est ainsi approché localement par un modèle simplifié.

La méthode LIME est l'une des premières méthodes d'interprétation locales de modèles et elle est largement utilisée aujourd'hui.

4.5.2 Principe de la méthode LIME

La méthode LIME a été introduite dans l'article *Why Should I trust You ?*[9] de Ribeiro et al. (2016).

LIME consiste à utiliser un modèle de substitution interprétable (noté M_2) qui approche localement au mieux un modèle de *machine learning* plus complexe (noté M_1).

Pour cela des d'observations bruitées sont simulées à partir du jeu de données initial puis nous calculons leur prédiction avec le modèle M_1. Ces nouvelles observations vont ensuite être pondérées selon leur proximité avec le jeu de données initial.

A partir de ce jeu de données pondéré, nous construisons des prédictions avec le modèle M_2, généralement de type Lasso pour la régression ou un arbre de décision pour la classification.

Nous obtenons ainsi la fonction $\hat{g}$ associée au modèle M_2 en résolvant le problème d'optimisation suivant :

$$\hat{g} = \underset{g \in G}{\operatorname{argmin}} [L(f, g, \pi_x) + \Omega(g)]$$

avec J la fonction de coût, f la fonction associée au modèle M_1, g la fonction associée au modèle M_2 qui appartient à la classe de modèle G, Ω la complexité du modèle et π_x la mesure de proximité au voisinage de l'observation x considérée pour interpréter le modèle M_1.

On peut résumer la méthode LIME de la manière suivante :

- les variables explicatives initiales sont permutées plusieurs fois pour chaque prédiction à expliquer
- la prédiction du modèle M_1 est considérée pour chaque permutation
- une distance convertie en score de proximité est calculée entre chaque permutation et l'observation initiale
- on sélectionne les n variables explicatives qui expliquent le mieux les prédictions du modèle M_1 sur les données permutées
- un nouveau modèle, plus simple, M_2, qui conserve les n variables explicatives, est calibré sur les nouvelles données pondérées par leur proximité avec l'observation initiale
- les poids associés aux n variables explicatives du modèle M_2 sont extraits et sont utilisés pour expliquer le comportement local du modèle M_1

4.5.3 Application au modèle de cessation d'activité

Désormais nous sommes capables de mieux comprendre notre algorithme de forêt aléatoire avec la méthode LIME.

Pour cela nous appliquons la méthode LIME avec $n = 5$ sur deux individus de notre base de test, une entreprise qui a cessé son activité et une entreprise toujours en activité pour apprécier au mieux l'influence des variables explicatives sur la prédiction de cessation d'activité.

Nous obtenons alors les coefficients des variables explicatives par une régression Lasso pour chaque entreprise :

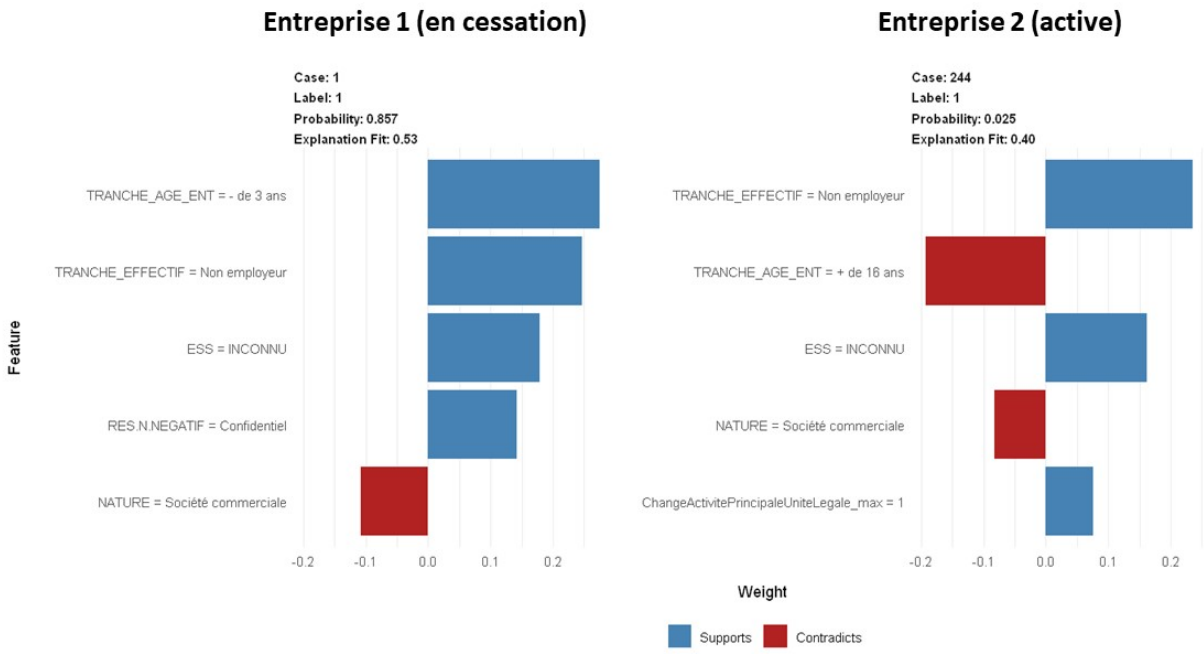

FIGURE 4.23 – Explication locale par méthode LIME sur deux entreprises

Nous pouvons tirer quelques conclusions à partir de ces graphiques.

Tout d'abord nous remarquons que les variables les plus influentes sur la cessation d'activité des entreprises 1 et 2 sont leur tranche âge, leur tranche d'effectif et leur appartenance à l'économie sociale et solidaire.

Dans le cas de l'entreprise 1, sa tranche d'âge (moins de 3 ans) et le fait qu'elle n'ait aucun employé contribuent positivement à sa prédiction de cessation d'activité vu que les coefficients associés à ces variables sont positifs.

Pour l'entreprise 2, sa tranche d'âge (+ de 16 ans) et sa nature juridique (société commerciale) contribuent à faire baisser la prédiction de cessation d'activité.

Ces conclusions sont bien cohérentes avec l'analyse univariée de la cessation d'activité partie 2.5.2.

Maintenant que nous avons sélectionné et interprété notre modèle de cessation d'activité nous allons désormais afficher quelques limites de cette modélisation.

4.6 Limites et conclusion de la modélisation de la cessation d'activité

Bien que notre modèle de cessation d'activité retenu affiche des bons critères de performance, une telle modélisation présente néanmoins certaines limites :

- la cessation d'activité est un phénomène complexe dont certaines causes n'ont pas pu être prises en compte dans la modélisation présentée dans ce mémoire comme le départ à la retraite ou le décès du dirigeant
- certaines données issues de l'Open Data sont parfois incomplètes ou confidentielles comme la base de données Chiffres Clés 2018 d'Infogreffe qui recense uniquement 1093562 entreprises

Pour conclure, maintenant que nous avons associé chaque entreprise du répertoire Sirene à un score de cessation d'activité, nous allons devoir le communiquer aux agents de Generali pour qu'ils pilotent mieux leur souscription de nouveaux contrats.

4.7 Communication du score de défaillance aux agents généraux

La dernière partie de ce chapitre consiste à détailler comment le score de cessation d'activité va être utilisé et mis à disposition de leurs utilisateurs finaux : les agents généraux de Generali France.

En effet il n'est pas pertinent de communiquer le score de cessation modélisé exact d'une entreprise aux agents car il n'est pas assez opérationnel et interprétable.

C'est pourquoi nous avons choisi de classer les entreprises du répertoire Sirene en 4 niveaux de risque de cessation : Peu Risqué, Risque Normal, Risqué et Très risqué. Ainsi nous associons une entreprise à un niveau de risque pour les agents.

Afin de déterminer les 4 niveaux, étudions d'abord la distribution du score de cessation par forêt aléatoire sur la base de test.

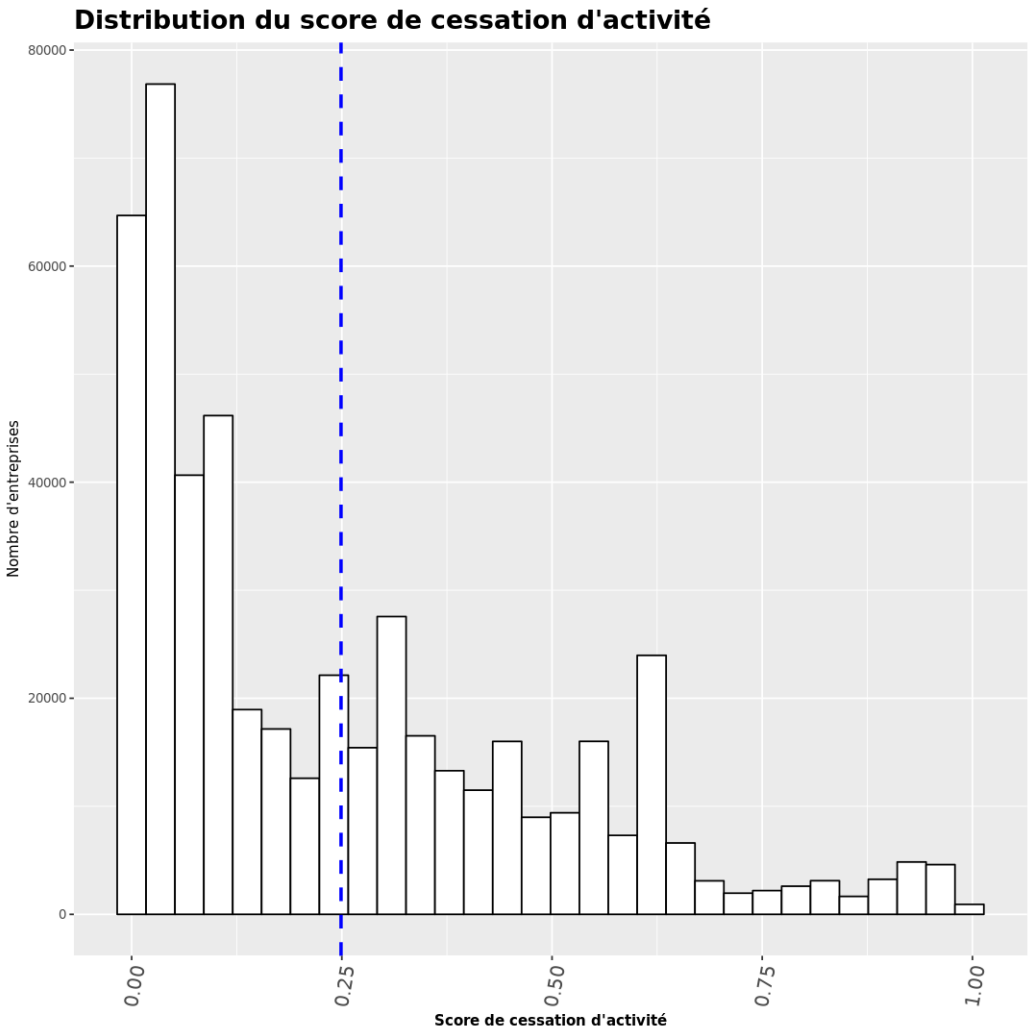

FIGURE 4.24 – Histogramme du score de cessation d'activité

En observant cet histogramme on constate que le modèle prédit en moyenne un score de cessation à 0.25 et qu'une majorité d'entreprises a un score inférieur. On s'attend donc à ce que plus le niveau de risque soit élevé, moins il contienne d'entreprises.

Pour déterminer les niveaux de risques nous avons calculé les taux de cessation réels en fonction des quantiles du score de cessation prédit, puis nous appliquons l'algorithme *k-means* avec 4 clusters qui seront les 4 niveaux de risque de cessation d'activité.

• **L'algorithme *k-means***

L'algorithme *k-means* est une méthode de classification non supervisée qui vise à partitionner un jeu de données en k clusters en minimisant la distance entre les points à l'intérieur de chaque cluster.

Afin de comparer le degré de similarité entre les différentes observations, on introduit la distance Euclidienne. Deux observations proches auront une distance petite, et inversement.

Soit une matrice X ayant n observations et p variables explicatives quantitatives. Dans l'espace vectoriel E^p, la distance Euclidienne entre deux observations x_1 et x_2 se calcule comme :

$$d(x_1, x_2) = \sqrt{\sum_{i=1}^{p}(x_{1,i} - x_{2,i})^2)}$$

Nous appliquons maintenant l'algorithme *k-means* sur nos données :

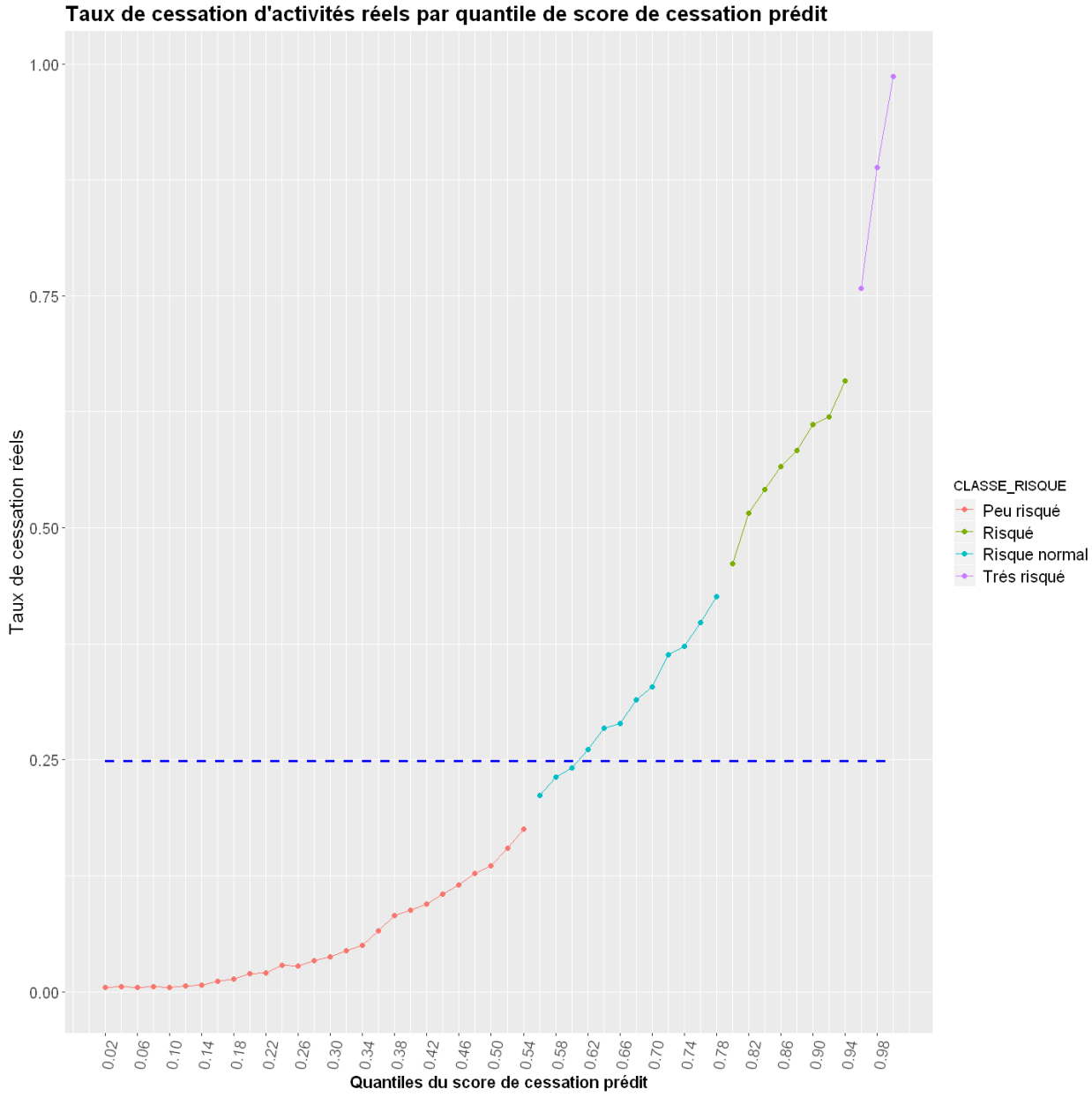

FIGURE 4.25 – Détermination des 4 niveaux de risque de cessation

Ainsi nous avons bien 4 niveaux de risque qui se dessinent en fonction du profil de risque de l'entreprise qui sont représentés dans le graphique suivant :

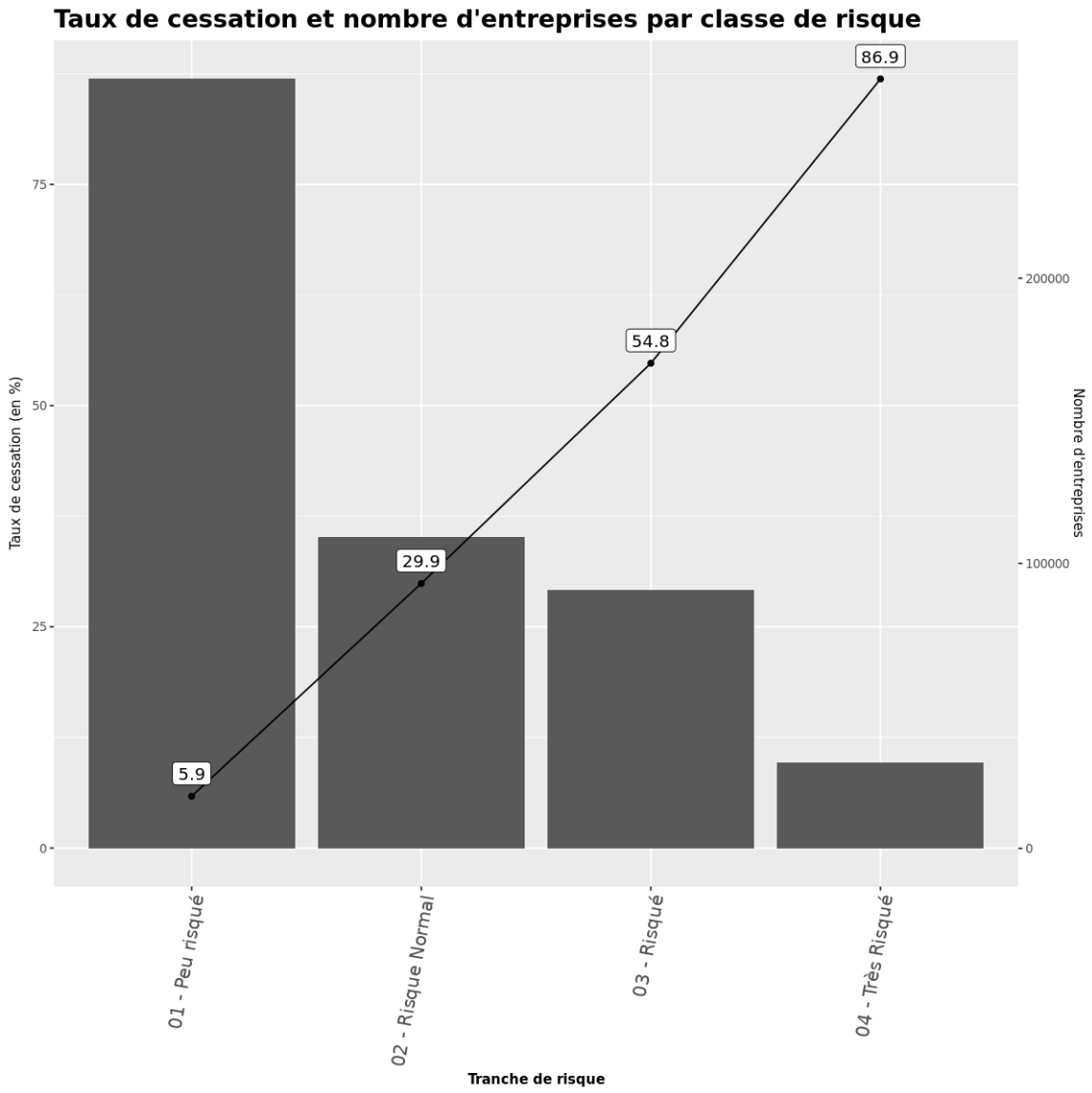

FIGURE 4.26 – Distribution des niveaux de risque avec leur taux de cessation moyen

Les entreprises catégorisées en "Peu risqué" ont un taux moyen de cessation d'activité de 5.9% contre 86.9% pour les entreprises catégorisées en "Très risqué".

Pour conclure nous recommandons aux agents de Generali France d'orienter la souscription de nouveaux contrats d'assurance aux entreprises "Peu risqué" et d'éviter de démarcher des entreprises catégorisées "Très risqué".

Troisième partie

Partie III : Application à la modélisation de la résiliation en MRC

Chapitre 5

Modélisation de l'acte de résiliation en MRC

5.1 Périmètre

La base de données contient tous les contrats actifs au 01/07/2017 associés aux produits 100% *Pro Artisans - Commerçants* et 100% *Pro services*.

La variable cible est *TOP_RESIL* qui vaut 0 si le contrat est encore actif au 01/07/2019, 1 sinon.

5.2 Données utilisées

Pour modéliser la résiliation en MRC nous utilisons les variables suivantes :

Données clients :

- Famille d'activité du client
- Est-que le client détient moins de contrat chez Generali au 01/07/2017 par rapport au 01/07/2016 ?
- Est-ce que le client détient plus d'un contrat chez Generali ?
- Est-ce que le client a des employés ?
- Région du client

Données contrats et risque :

- Type de prélèvement (annuel, mensuel ou semestriel)
- Ancienneté du contrat
- Tranche de prime du contrat
- Tranche du capital incendie assuré
- Type d'occupant (locataire, propriétaire)
- Réseau de distribution attaché au contrat (exemple : agent, courtier)

Les différentes modalités de ces variables sont renseignées dans l'annexe 4 du mémoire.

La base finale contient 42 182 contrats pour 10 434 résiliations, soit un taux de résiliation de 24.7%

5.3 Analyse descriptive de la résiliation

Dans cette section nous allons analyser l'évolution du taux de résiliation suivant quelques variables explicatives pour mieux comprendre le phénomène de résiliation.

Ancienneté du contrat

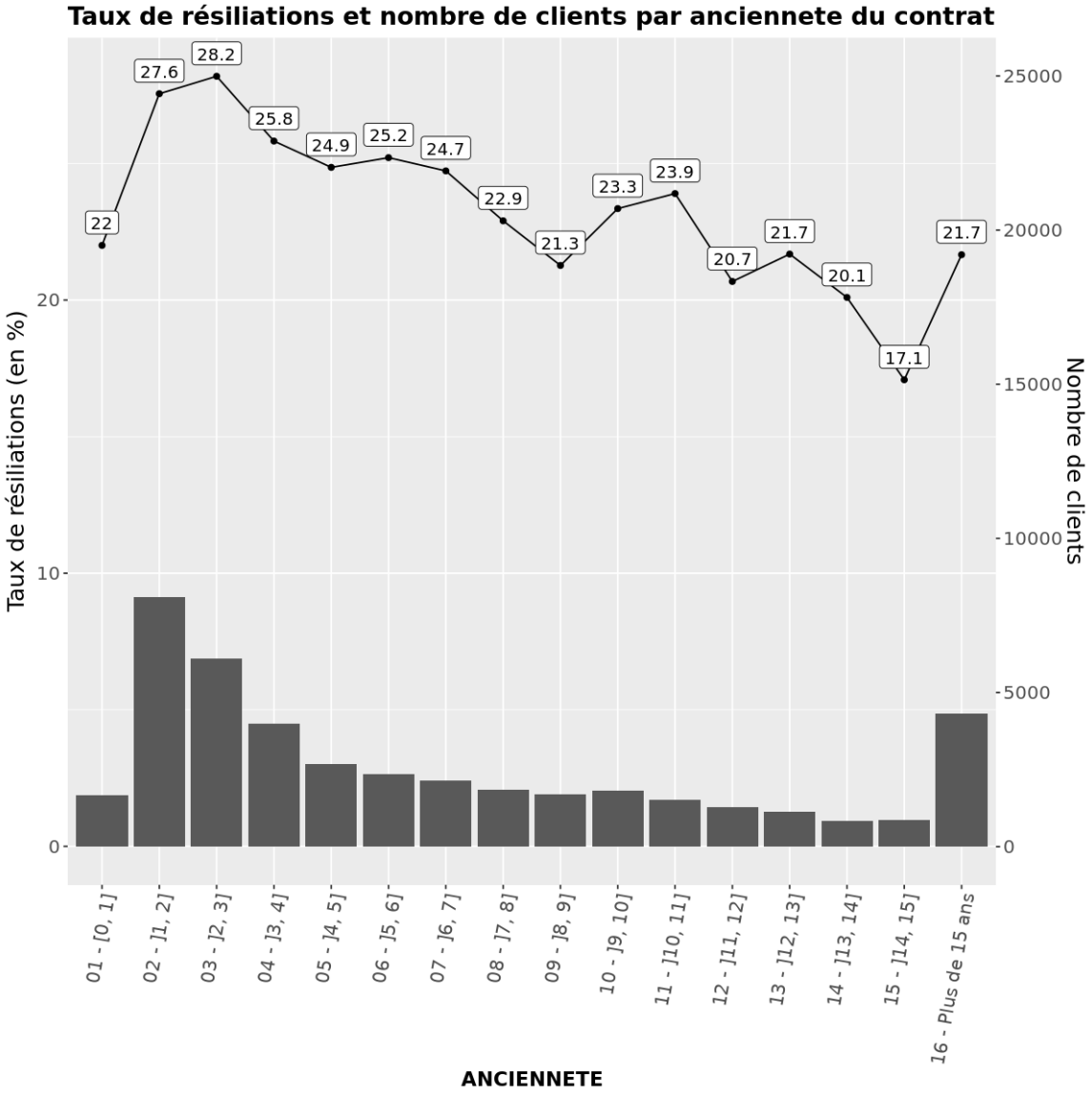

FIGURE 5.1 – Taux de résiliation par ancienneté du contrat

Le taux de résiliation moyen décroît en fonction de l'ancienneté du contrat avant de remonter de nouveau pour les contrats âgés de plus de 15 ans.

Détention du client

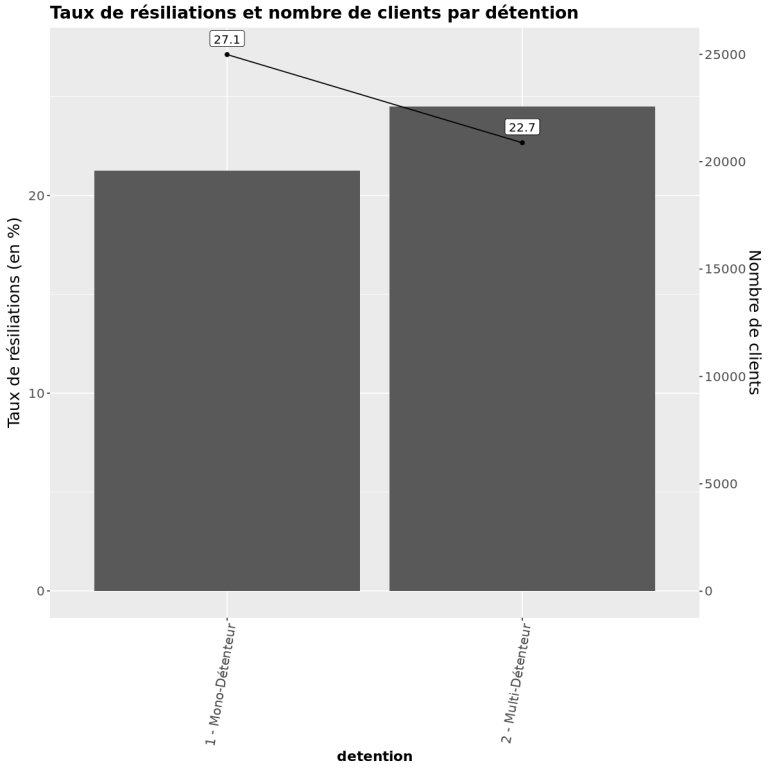

FIGURE 5.2 – Taux de résiliation par détention du client

A partir de ce graphique nous observons qu'en moyenne les clients multi-équipés ont tendance à moins résilier leur contrat MRC que les autres.

5.4 Ajout du niveau de risque de cessation d'activité

Afin d'enrichir notre base clients du score de cessation d'activité, nous devons isoler les clients qui ont un numéro siren pour compléter la base clients avec la base Sirene et ainsi récupérer les variables explicatives du modèle de cessation d'activité.

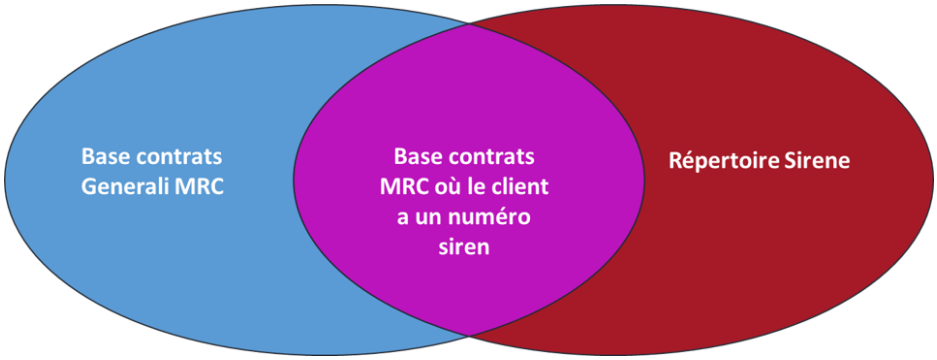

FIGURE 5.3 – Population d'étude pour modéliser la résiliation en MRC

Ainsi nous allons modéliser la cessation d'activité des clients qui possèdent un contrat MRC et un numéro siren.

Par ailleurs nous tenons compte que cette population d'étude présente un taux de cessation d'activité de $4,6\%$ contre $23,2\%$ dans le répertoire Sirene : c'est à dire qu'entre le 01/07/2017 et le 01/07/2019 $4,6\%$ des clients Generali avec un contrat MRC et un numéro siren ont cessé leur activité.

Pour modéliser la cessation d'activité, il suffit d'utiliser le meilleur modèle de cessation d'activité choisi dans la Partie II du mémoire, la forêt aléatoire.

Une fois le score de cessation rajouté à la base clients, nous pouvons étudier s'il y a une liaison entre le score de cessation et la résiliation.

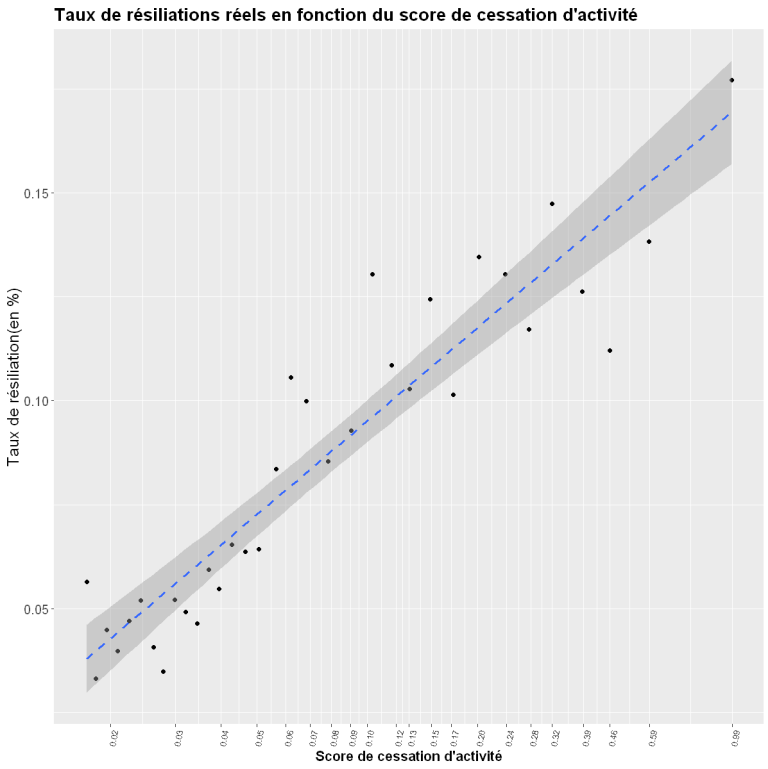

FIGURE 5.4 – Taux de résiliations en fonction du score de cessation

D'après le graphique, une tendance linéaire semble exister entre le taux de résiliation et le score de cessation. C'est pourquoi comme précédemment nous pouvons regrouper les clients MRC par niveau de risque de cessation d'activité avec les mêmes seuils que la Partie II.

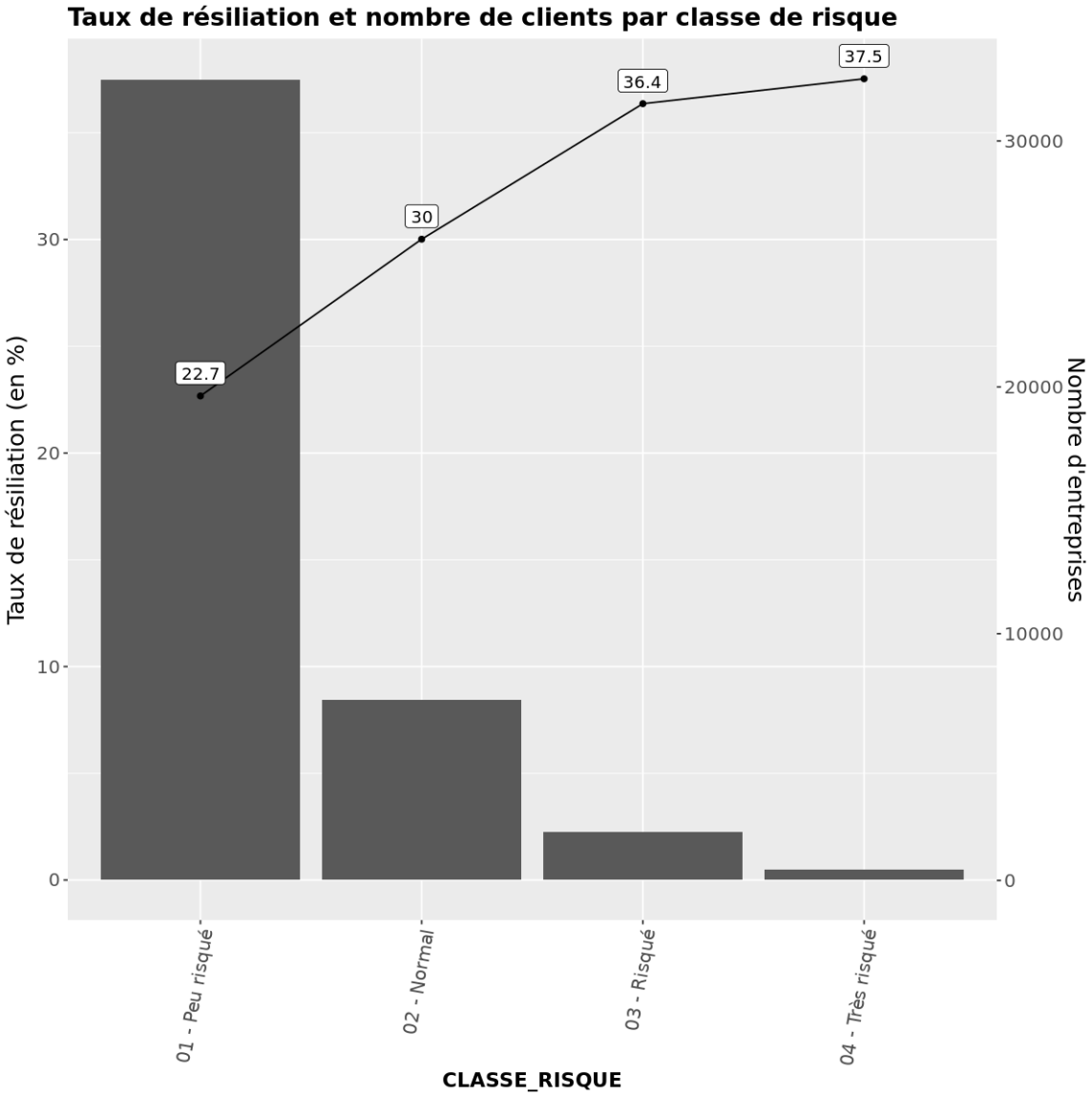

FIGURE 5.5 – Taux de résiliation par niveau de risque de cessation

Comme attendu, le taux moyen de résiliation est d'autant plus élevé que l'on se trouve dans un niveau de risque de cessation d'activité élevée.

Cette première étude descriptive nous montre qu'il semble exister un lien non négligeable entre notre score de cessation d'activité et la résiliation en MRC. Le but de la section suivante et de vérifier statistiquement cette liaison.

5.5 Analyse statistique du niveau de risque de cessation sur la résiliation

5.5.1 Test d'indépendance du χ^2

Le test d'indépendance du Khi-Deux permet de vérifier l'absence de lien statistique entre deux variables X et Y. Les deux sont dites indépendantes lorsqu'il n'existe aucun lien statistique entre elles.

L'hypothèse nulle $(H0)$ de ce test est la suivante : les deux variables X et Y sont indépendantes.

Le test d'indépendance du χ^2 est ensuite basé sur la statistique suivante avec T l'effectif théorique et O l'effectif observé :

$$D^2 = \sum_{i,j} \frac{(O_{i,j} - T_{i,j})^2}{T_{i,j}}$$

La loi de T suit asymptotiquement une loi du χ^2 à $(I-1)(J-1)$ degrés de liberté.

Dans notre cas, la fonction *chisq.test* du logiciel R nous donne une p-value $\leq \alpha = 5\%$.

```
        Pearson's Chi-squared test

data:  table(BASE_TRAVAIL_MRC$CLASSE_RISQUE, BASE_TRAVAIL_MRC$TOP_RESIL)
X-squared = 204.11, df = 3, p-value < 0.00000000000000022
```

FIGURE 5.6 – Test d'indépendance du Khi-Deux entre le niveau de risque de cessation d'activité et l'acte de résiliation en MRC

Nous rejetons donc l'hypothèse d'indépendance au niveau de risque $\alpha = 5\%$.

Notre intuition précédente est confirmée : il existe bien un lien statistique entre le niveau de risque de cessation et la résiliation.

5.5.2 Étude des liaisons de la base MRC

Liaisons entre les variables explicatives

Les variables explicatives de la base MRC sont qualitatives ou quantitatives discrétisées, nous utilisons le V de Cramer comme mesure d'association entre elles.

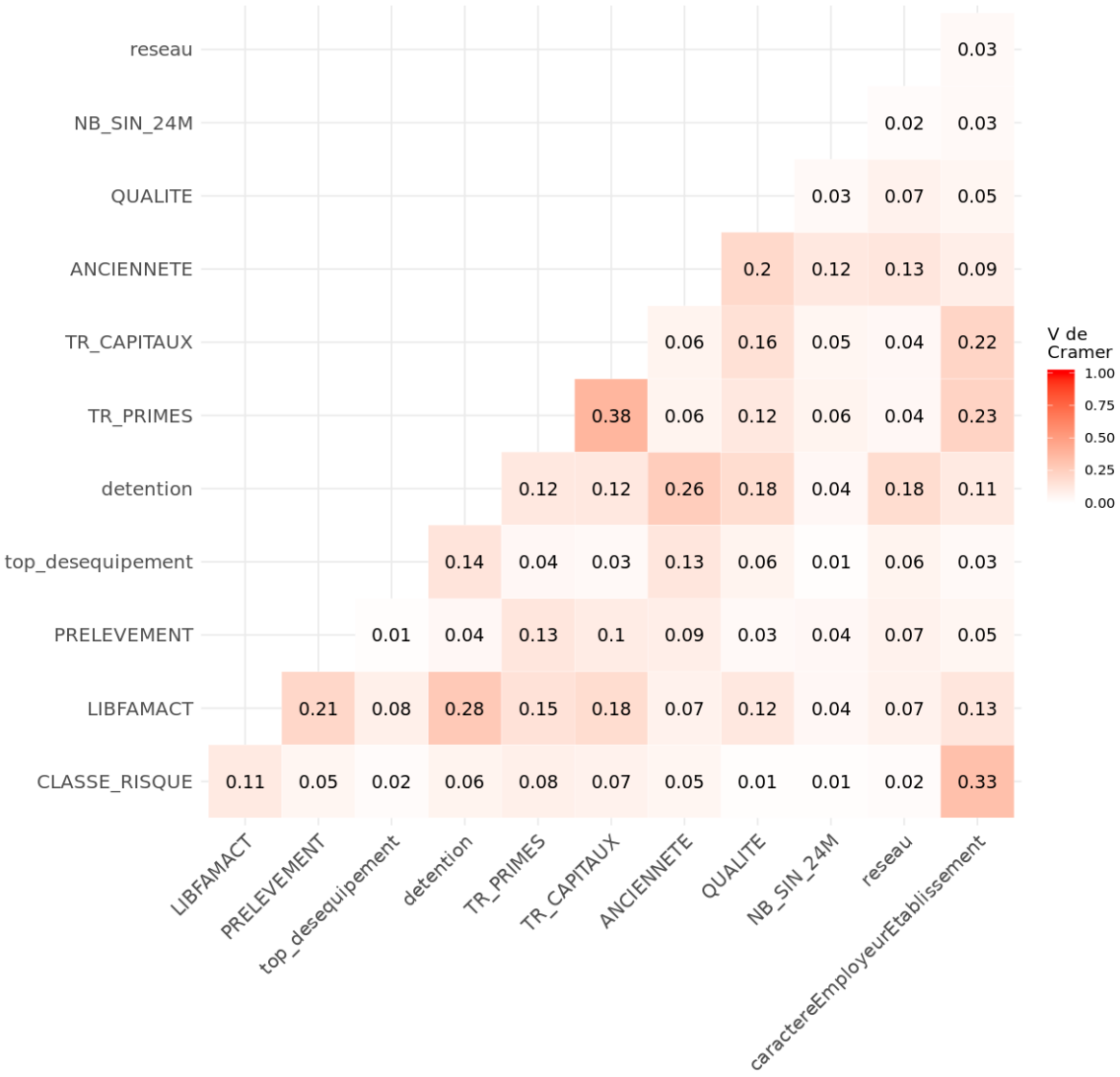

FIGURE 5.7 – V de Cramer entre les variables explicatives de la base MRC

Cette matrice d'association nous indique que notre nouvelle variable créée dans la partie II *CLASSE_RISQUE* est très peu corrélée aux autres variables de la base MRC. Elle apporte donc de nouvelles informations et nous pouvons l'intégrer dans un modèle type GLM.

Liaisons avec la variable cible

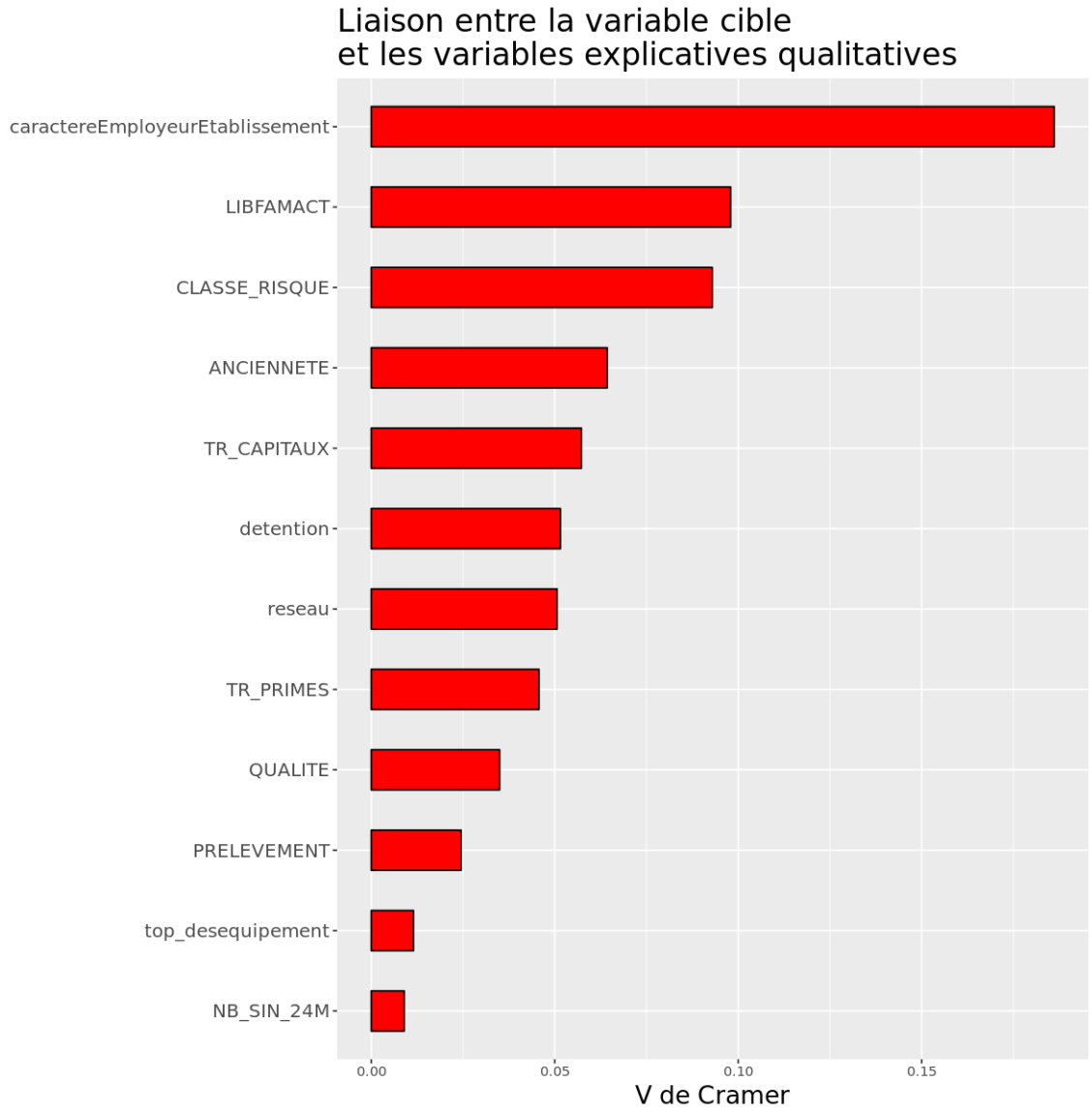

FIGURE 5.8 – V de Cramer entre les variables explicatives de la base MRC et la variable cible

La variable *CLASSE_RISQUE* possède la deuxième liaison la plus élevée avec la variable cible derrière la famille d'activité du client *LIBFAMACT*.

Maintenant que l'on a vérifié que la nouvelle variable *CLASSE_RISQUE* est statistiquement liée à la variable cible et qu'elle apporte des nouvelles informations en plus de la base MRC initiale, nous allons vérifier si elle permet d'améliorer ou non le modèle de résiliation MRC.

5.6 Modélisation de la résiliation en MRC

Pour modéliser la résiliation en MRC et pouvoir mesurer l'impact de la variable *CLASSE_RISQUE*, nous allons utiliser le modèle de régression logistique qui est plus interprétable que les autres algorithmes utilisés dans la partie II du mémoire.

Ainsi nous allons modéliser la résiliation en MRC avec deux modèles de régression logistique : un avec la variable explicative *CLASSE_RISQUE* et un sans, puis nous allons comparer leur performance et mesurer l'impact de cette nouvelle variable.

Une sélection de variables stepwise est effectuée avant modélisation.

5.6.1 Modélisation sans la classe de risque

Test de l'hypothèse nulle globale : BETA=0			
Test	Khi-2	DDL	Pr > Khi-2
Likelihood Ratio	1 633,07	89	<,0001
Score	1 664,72	89	<,0001
Wald	1 572,70	89	<,0001

Analyse des effets type 3			
Effect	DDL	Khi-2 de Wald	Pr > Khi-2
caractereEmployeurEtablissement	1	987,51	<,0001
LIBFAMACT	37	234,02	<,0001
ANCIENNETE	15	88,93	<,0001
reseau	5	63,19	<,0001
LIBELLE_REGION	14	40,99	0,0002
TR_CAPITAUX	9	26,67	0,0016
QUALITE	3	21,93	<,0001
top_desequipement	1	18,50	<,0001
PRELEVEMENT	3	9,84	0,02
detention	1	7,23	0,0072

FIGURE 5.9 – Tests statistiques de la régression logistique sans la classe de risque

Toutes les variables sont significatives au seuil α = 5%. L'AIC de ce modèle est de 36 208.

5.6.2 Modélisation avec la classe de risque

Test de l'hypothèse nulle globale : BETA=0			
Test	Khi-2	DDL	Pr > Khi-2
Likelihood Ratio	1 675,83	92	<,0001
Score	1 710,77	92	<,0001
Wald	1 611,87	92	<,0001

Analyse des effets type 3			
Effect	DDL	Khi-2 de Wald	Pr > Khi-2
caractereEmployeurEtablissement	1	808,33	<,0001
LIBFAMACT	37	224,85	<,0001
ANCIENNETE	15	87,68	<,0001
reseau	5	64,45	<,0001
CLASSE_RISQUE	3	44,02	<,0001
LIBELLE_REGION	14	40,42	0,0002
TR_CAPITAUX	9	27,67	0,0011
QUALITE	3	23,53	<,0001
top_desequipement	1	17,80	<,0001
PRELEVEMENT	3	9,10	0,028
detention	1	6,87	0,0087

FIGURE 5.10 – Tests statistiques de la régression logistique avec la classe de risque

Toutes les variables sont significatives au seuil α = 5% y compris notre nouvelle variable *CLASSE_RISQUE* qui est la $5^{ème}$ variable la plus significative du modèle

parmi les 11 variables explicatives. L'AIC de ce modèle est de 36 171, très proche de l'AIC du modèle précédent.

On peut maintenant analyser les odds-ratio de la variable *CLASSE_RISQUE* pour comprendre l'influence de chaque niveau de risque sur la résiliation "toutes choses égales par ailleurs" .

L'analyse "toutes choses égales par ailleurs" consiste à mesurer l'effet propre d'un facteur sur le phénomène étudié en choisissant une modalité de référence à laquelle vont être comparées les autres modalités du facteur. Nous choisissions la modalité "01 - Peu risqué" comme modalité de référence pour la variable *CLASSE_RISQUE*.

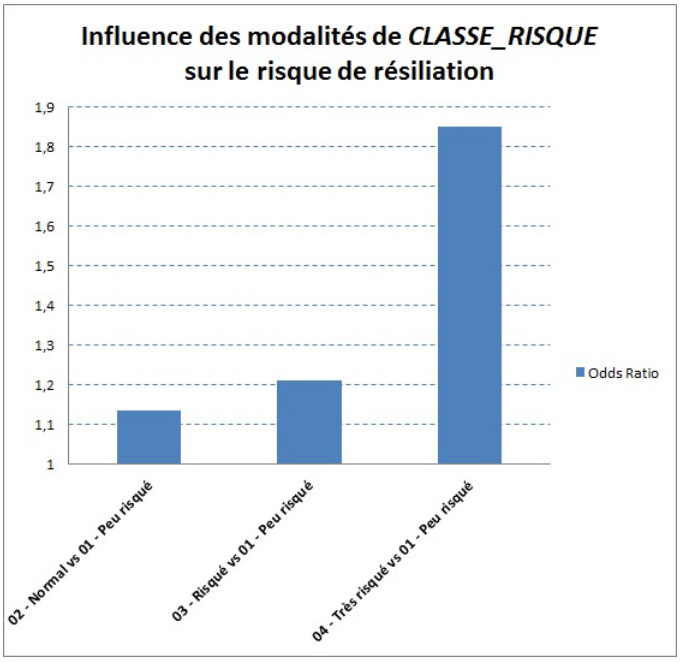

FIGURE 5.11 – Odds ratio de la régression logistique

Au vu du graphique, tous les odds de la variable *CLASSE_RISQUE* sont strictement supérieurs à 1 et ils sont croissants suivant les modalités.

Par exemple, toutes choses égales par ailleurs, une entreprise "très risquée" au sens de la cessation d'activité accuse 1.85 fois plus de chances de résilier un contrat MRC qu'une entreprise "Peu risqué".

Ces résultats sont cohérents avec l'étude descriptive précédente.

5.6.3 Comparaison des performances des modèles de résiliation

Les performances des deux modèles sont évaluées sur la base d'apprentissage et la base de test.

Modèle	Base	AUC	Lift à 20%	Précision pondérée	AIC
Avec classe de risque	Train	0,643	0,329	0,604	36 171
	Test	0,629	0,325	0,597	
Sans classe de risque	Train	0,641	0,328	0,605	36 208
	Test	0,629	0,324	0,595	

FIGURE 5.12 – Performances des deux modèles de résiliation

Au vu des performances des deux modèles de résiliation, nous observons que le modèle intégrant le risque de cessation est équivalent au modèle qui n'en prend pas compte : l'ajout du score de cessation n'améliore pas le pouvoir prédictif du modèle de résiliation.

Par ailleurs, nous avons refait la même démarche en gardant uniquement les résiliations causées par une disparition du risque. Cependant les performances ne sont toujours pas améliorées.

5.7 Conclusion sur l'application du score de cessation à la modélisation de la résiliation en MRC

Nous concluons que modéliser la résiliation en MRC en ajoutant le score de cessation d'activité comme variable explicative ne permet pas d'améliorer le pouvoir prédictif du modèle.

Pour en comprendre les raisons, nous pourrions aussi ré-échantillonner la base MRC de sorte à avoir autant d'individus dans chaque niveau de risque puis modéliser de nouveau la résiliation sur cette nouvelle base et ainsi mieux comprendre son influence niveau par niveau.

Néanmoins cette nouvelle variable pourra tout de même être exploitée dans d'autres études comme la modélisation du score de fragilité sur le périmètre des professionnels et des entreprises ou dans d'autres modèles de résiliation sur d'autres produits.

Conclusion générale

Ce mémoire a deux objectifs :

- Modéliser la cessation d'entreprise avec des données issues de l'Open Data pour accompagner les agents généraux de Generali dans leur souscription en leur indiquant les prospects professionnels à fort risque de cessation. Cette modélisation est faite en appliquant des modèles classiques comme un GLM ou un arbre CART et des méthodes d'apprentissage statistique comme l'algorithme XGBoost ou les forêts aléatoires
- Essayer d'améliorer un modèle de résiliation en Assurance MultiRisques Commerce en intégrant le niveau de risque de cessation d'activité.

Tout d'abord, il est difficile de modéliser la cessation d'activité car c'est un phénomène complexe dont certaines causes ne peuvent pas être intégrées dans un modèle.

Par exemple les données disponibles en Open Data ne permettent pas de capter certains facteurs de cessation d'activité comme le départ des fondateurs en retraite ou à l'étranger.

Néanmoins le modèle retenu pour modéliser la cessation d'activité affiche de bons indicateurs de performance grâce aux puissants algorithmes d'apprentissage.

C'est d'autant plus concluant que dans ce mémoire nous recherchons le modèle au meilleur pouvoir prédictif en mettant de côté son interprétabilité.

Nous avons ainsi construit quatre niveaux de risque de cessation d'activité qui seront communiqués aux agents généraux de Generali France pour les accompagner dans leur démarchage commerciale.

Par la suite nous avons voulu appliquer ce score de cessation d'activité à la résiliation de contrats d'assurance.

En effet, après avoir constaté le lien entre ces deux phénomènes, il semblait pertinent de vérifier si un modèle de résiliation pouvait être amélioré en intégrant le niveau de risque de cessation d'activité.

Ainsi nous avons choisi le produit d'Assurance MultiRisques Commerce (MRC) pour vérifier cette hypothèse.

Nous avons d'abord constaté une relation entre le niveau de risque de cessation d'activité et le taux moyen de résiliation : plus une entreprise a de chances de cesser son activité, plus elle a de chances de résilier son contrat MRC.

Ensuite pour mesurer l'influence du risque de cessation d'activité sur le risque de résiliation, nous avons modélisé l'acte de résiliation avec deux modèles de régression logistique : un qui intègre le niveau de risque de cessation du client, l'autre non.

Malgré la relation a priori entre ces deux phénomènes, nous avons obtenu des résultats en dessous de nos attentes : les performances du modèles qui intègrent le niveau de risque sont équivalentes au modèle initial.

Nous avons présenté plusieurs pistes pour mieux en comprendre les raisons.

Ainsi, bien que le score de cessation d'entreprise soit un bon indicateur individuellement, l'utilisation de ce score comme variable explicative dans le modèle de résiliation MRC n'a pas été concluante.

En revanche le score de cessation pourra dorénavant faire partie des variables explicatives que nous testerons dans nos prochains scores de résiliation sur le périmètre des professionnels et des entreprises, ou dans d'autres scores sur les clients de Generali.

Pour conclure, le premier objectif qui répond au besoin initial exprimé par Generali est atteint. Quant au second, les premiers résultats ne sont pas encore concluants mais le score de cessation est désormais une nouvelle variable que nous pourrons utiliser pour d'autres études.

Annexes

Annexe 1 : Source des jeux de données utilisées

Pour être en conformité avec la « Licence Ouverte / Open Licence » :

Titre du jeu de données	Producteur	Dernière date de mise à jour	Lien des données originales
Base Sirene des entreprises et de leurs établissements (SIREN, SIRET)	Insee	6 septembre 2019	https://www.data.gouv.fr/fr/datasets/base-sirene-des-entreprises-et-de-leurs-etablissements-siren-siret/
BODACC	Premier ministre	16 septembre 2019	https://www.data.gouv.fr/fr/datasets/bodacc/
Chiffres Clés 2018	Infogreffe	12 juillet 2019	https://www.data.gouv.fr/fr/datasets/chiffres-cles-2018/

Annexe 2 : Variables explicatives retenues pour modéliser la cessation d'activité

Variable	Modalité
TRANCHE_EFFECTIF	0 salarié
	1 ou 2 salariés
	3 salariés ou +
	Effectif Inconnu
	Non employeur
CATEGORIE	ETI
	GE
	PME
caractereEmployeurEtablissement	O
	N
NATURE	Artisan
	Artisan-commerçant
	Autre personne morale immatriculée au RCS
	Commerçant
	Groupement de droit privé
	Personne morale de droit étranger
	Société commerciale
LIBELLE_REGION	Auvergne-Rhône-Alpes
	Bourgogne-Franche-Comté
	Bretagne
	Centre-Val de Loire
	Corse
	Grand Est
	Hauts-de-France
	Île-de-France
	Inconnu
	Normandie
	Nouvelle-Aquitaine
	Occitanie
	Outre-Mer
	Pays de la Loire
	Provence-Alpes-Côte d'Azur
CATEGORIE_COMMUNE	Autre commune multipolarisée
	Commune appartenant à la couronne d'un grand pôle
	Commune appartenant à la couronne d'un moyen pôle
	Commune appartenant à la couronne d'un petit pôle
	Commune appartenant à un grand pôle (10 000 emplois ou plus)
	Commune appartenant à un moyen pôle (5 000 à moins de 10 000 emplois)
	Commune appartenant à un petit pôle (de 1 500 à moins de 5 000 emplois)

CATEGORIE_COMMUNE	Commune isolée hors influence des pôles
	Commune multipolarisée des grandes aires urbaines
	Inconnu
ChangeNicSiegeUniteLegale_max	0
	1
ChangeActivitePrincipaleUnite Legale_max	0
	1
ChangeCaractereEmployeurUnite Legale_max	0
	1
ChangeNom	0
	1
TRANCHE_AGE_ENT	- de 3 ans
	3 à 5 ans
	6 à 10 ans
	11 à 15 ans
	+ de 16 ans
ESS	INCONNU
	N
	O
	Pas de ESS
ACTIVITE_ENTREPRISE	ACTIVITÉS FINANCIÈRES ET D'ASSURANCE
	ACTIVITÉS IMMOBILIÈRES
	AGRICULTURE, SYLVICULTURE ET PÊCHE
	AUTRES ACTIVITÉS DE SERVICES
	COMMERCE ET RÉPARATION AUTOMOBILE
	CONSTRUCTION
	ENSEIGNEMENT, SANTÉ HUMAINE, ACTION SOCIALE ET SERVICES AUX MÉNAGES
	HÉBERGEMENT ET RESTAURATION
	INDUSTRIE
	INFORMATION ET COMMUNICATION
	SOUTIEN AUX ENTREPRISES
	TRANSPORTS ET ENTREPOSAGE
FAMILLE_JUGEMENT	Arrêt de la Cour d'Appel
	Avis de dépôt
	Extrait de jugement
	Jugement d'ouverture
	Jugement de clôture
	Jugement prononçant
	Loi de 1967
	Rétractation sur tierce opposition
	INCONNU
RES.N.NEGATIF	Confidentiel
	FALSE
	TRUE
TRANCHE_CA_1	[-8.8e+06,2.08e+05]
	(2.08e+05,3.74e+06]
	(3.74e+06,6.87e+07]
	(6.87e+07,5.74e+10]

Annexe 3 : Méthode SHAP

L'objectif principal de la méthode SHAP est de calculer la contribution de chaque variable dans la prédiction faîte par une modèle, en leur attribuant un score dit "juste".

Pour cela, la méthode SHAP repose sur le calcul des valeurs de Shapley, issues de la théorie des jeux.

Pour comprendre le fonctionnement de cette méthode, rappelons d'abord le principe de la valeur de Shapley dans la théorie des jeux.

Valeur de Shapley en théorie des jeux

Nous nous plaçons dans le domaine de la théorie des jeux et plus particulièrement en jeu coopératif pour introduire la valeur de Shapley.

Soit un jeu coopératif donné par un ensemble de joueurs $P = \{1, 2, ..., p\}$ et v la fonction caractéristique du gain de tout sous-ensemble de joueurs $S \in S(P)$.

Notons $\phi_i(v)$ le montant réparti pour chaque participant $i \in P$.

Alors la solution unique et équitable d'une répartition de ces montants est l'unique solution des axiomes suivants selon Lloyd Shapley en 1953 :

- Efficacité : $\sum_{i=1}^{p} \phi_i(v) = v(P)$: la somme des montants attribués aux joueurs de P doit être égale au profit global obtenu par la coalition des joueurs
- Symétrie : pour tout couple de joueurs $(i, j) \in \{1, ..., p\}^2$, si $\forall S \in S(P \backslash \{i, j\})$, $v(S \cup i) = v(i \cup S)$, alors $\phi_i(v) = \phi_j(v)$: : si deux joueurs i et j peuvent se substituer dans une sous-coalition S de P alors ils perçoivent les mêmes montants
- Le joueur nul : soit i un joueur et S une coalition, on a : $v(S \cup i) = v(S) \Rightarrow i$ est nul pour v et $\phi_i(v) = 0$
- Additivité : soit i un individu participant à deux jeux composés des mêmes joueurs dont les fonctions caractéristiques pour chacun de ces jeux sont v et w. On a alors la relation suivante : $\phi_i(v) + \phi_i(w) = \phi_i(u + w)$

La valeur de Shapley, solution unique et équitable du gain total du joueur i est donnée par :

$$\phi_i(v) = \sum_{S \in S(P) \backslash \{i\}} \frac{(p - |S| - 1)!S!}{p!} (v(S \cup i) - v(S))$$

Soit j une variable de valeur $y_{i,j}$ associée à un individu i. La valeur de Shapley peut alors être interprétée comme la contribution marginale moyenne de la valeur $y_{i,j}$ de la variable j. D'où :

$$\phi_{i,j}(v) = \sum_{S \in S(P) \backslash \{i\}} (v(S) - v(S \backslash i))$$

Estimation de la Valeur de Shapley par Monte-Carlo

On peut approximer la Valeur de Shapley par Monte-Carlo de la façon suivante en considérant la contribution de toute variable j sur la prédiction d'un individu x_i :

$$\hat{\phi_{i,j}}(v) = \frac{1}{M} \sum_{k=1}^{M} (\phi_{i,j}(v)^k)$$

Annexe 4 : Variables explicatives pour modéliser la résiliation en MRC

Variable	Description	Modalité
LIBFAMACT	Secteur d'activité du client	Bar restaurant
		Bureaux seuls
		Activites Medicales - Paramedicales
		Habillement mode et accessoire
		Services au particulier et/ou professionnel et entreprise
		Papier Impression
		Alimentaire
		Beaute Hygiene Esthetique
		Soins aux animaux
		Hotel hebergement
		Sante
		Juridique
		Commerces et services divers
		Metal
		Informatique Electromenager Telephonie
		Loisirs musique
		Batiment
		Autos cycles bateaux aeronautique
		Floral et Animaux
		Amenagement decoration
		Bricolage
		Bois
		Comptabilite - Finance
		Metiers de l'audiovisuel et de l'écriture
		SSII
		Conseil en entreprise
		Evenementiels, Spectacles
		Verre
		Electronique electricite
		Services a la personne
		Plastique
		Expertise
		Enseignement
		Bureaux Etudes - Controles
		Pierre
		Activites de Bien-etre
		Sport
PRELEVEMENT	Type de prélèvement de la prime	MENSUEL
		ANNUEL
		SEMESTRIEL
		TRIMESTRIEL
top_desequipement	Est-ce que le client détient moins de contrats chez Generali au 31/07/2017 par rapport au 31/07/2016 ?	N
		O
detention	Critère de multi-équipement du client	1 - Mono-Détenteur
		2 - Multi-Détenteur
TR_PRIMES	Tranche de la prime du contrat MRC	01 - Moins de 100€
		02 - 100€ à 200€
		03 - 200€ à 300€
		04 - 300€ à 400€
		05 - 400€ à 500€
		06 - 500€ à 600€
		07 - 600€ à 700€
		08 - 700€ à 800€
		09 - 800€ à 900€
		10 - 900€ à 1.000€
		11 - 1.000€ à 1.250€
		12 - 1.250€ à 1.500€
		13 - 1.500€ à 2.000€
		14 - 2.000€ à 2.500€
		15 - 2.500€ à 3.000€
		16 - 3.000€ à 3.500€
		17 - 3.500€ à 4.000€
		18 - Plus de 4.000€

TR_CAPITAUX	Tranche du capital incendie assuré	01 - Moins de 5.000€
		02 - 5.000€ à 10.000€
		03 - 10.000€ à 15.000€
		04 - 15.000€ à 25.000€
		05 - 25.000€ à 35.000€
		06 - 35.000€ à 50.000€
		07 - 50.000€ à 70.000€
		08 - 70.000€ à 110.000€
		09 - 110.000€ à 210.000€
		10 - + de 210.000€
ANCIENNETE	Ancienneté du contrat en années	01 - [0, 1]
		02 -]1, 2]
		03 -]2, 3]
		04 -]3, 4]
		05 -]4, 5]
		06 -]5, 6]
		07 -]6, 7]
		08 -]7, 8]
		09 -]8, 9]
		10 -]9, 10]
		11 -]10, 11]
		12 -]11, 12]
		13 -]12, 13]
		14 -]13, 14]
		15 -]14, 15]
		16 - Plus de 15 ans
QUALITE	Qualité de l'occupant	Locataire
		Propriétaire
		Copropriétaire
		NR
reseau	Réseau de distribution associé au contrat	Agent Signataire
		Agent Non Signataire
		Courtier Premium
		Courtier Expert
		Courtier Affaire
		Courtier Classique

Bibliographie

[1] R. Bellina. *Méthodes d'apprentissage appliquées à la tarification non-vie.* Mémoire ISFA, 2014.

[2] L. Breiman. *Random Forests.* Machine Learning, 45, 5–32, 2001.

[3] L. Breiman, J. Friedman, C. J. Stone, and R.A. Olshen. *Classification and Regression Trees.* Mathematics, 1984.

[4] V. Carrasco. *Les caractéristiques des entreprises qui font l'objet d'une ouverture de procédure collective – 2006-2012.* Ministère de la Justice – SDSE, 2014.

[5] T. Chen and C.Guestrin. *XGBoost : A Scalable Tree Boosting System.* 2016.

[6] B. De Moura Fernandes. *Défaillances d'entreprises en France : bilan 2018 et perspectives 2019.* Coface, 2019.

[7] J. H. Friedman. *Greedy function approximation : A gradient boosting machine.* The Annals of Statistics, 2001.

[8] R. Genuer and J-M Poggi. *Arbres CART et Forêts aléatoires, Importance et sélection de variables.* 2017.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin. *"Why Should I Trust You ?" - Explaining the Predictions of Any Classifier.* arXiv :1602.04938„ 2016.

[10] J. Rodriguez. *Interpretability vs. Accuracy : The Friction that Defines Deep Learning.* `https://towardsdatascience.com/interpretability-vs-accuracy-the-friction-that-defines-deep-learning-dae16c84db5c`, 2018.

[11] L. Rouviere. *Introduction aux méthodes d'agrégation : boosting, bagging et forêts aléatoires. Illustrations avec R.* 2015.