

MÉMOIRE DE RECHERCHE



**PREVISION DES DÉFAILLANCES
DES ENTREPRISES AVEC LA
BIBLIOTHEQUE DE MACHINE
LEARNING XGBOOST**

PRÉSENTÉ PAR : MARIE-CLAIRE FÉVRIER


**MINISTÈRE
DES ARMÉES**
*Liberté
Égalité
Fraternité*



**AGENCE
INNOVATION
DÉFENSE**

Année académique 2020 – 2021

Mémoire de recherche

pour obtenir le titre de

Mastère 2 in Artificial Intelligence & Management
délivré par l'IA School

présenté par

Marie-Claire Février

Classe M2C, parcours Data Scientist

Le 7 septembre 2021

**Prévision des défaillances des entreprises avec la bibliothèque
de Machine Learning XGBoost**

Entreprise d'accueil : Ministère des Armées

Tuteurs : Michaël Krajecki, Mattis Paulin

Jury :

Table des matières

Résumé.....	8
Abstract	9
Remerciements	12
1 Introduction.....	14
2 Contexte du projet de prédiction des défaillances	27
3 Revue de littérature	28
3.1 Cadre théorique de la défaillance d'entreprise	28
3.1.1 Notion de défaillance d'entreprise	28
3.1.1.1 Différents évènements de la vie d'une entreprise.....	28
3.1.1.2 Hypothèse de l'impact des procédures amiables sur la production industrielle.....	29
3.1.1.3 Cessation d'activité d'une entreprise.....	30
3.1.1.4 Caractérisation de la défaillance dans les thèses précédentes.....	30
3.1.2 Eventuels facteurs explicatifs de la défaillance	31
3.1.3 « Trous de connaissances » en prédiction de défaillance.....	33
3.1.4 Aspects « Métier » de la problématique de recherche	33
3.2 Cadre théorique de la prédiction en classification	34
3.2.1 Etat de l'art en classification	34
3.2.2 Classification dans un contexte de déséquilibre de classe	36
3.2.2.1 Affichage des résultats de la classification avec la matrice de confusion. 36	
3.2.2.2 Mesure du déséquilibre avec le Ratio de déséquilibre	37
3.2.2.3 Problèmes posés pour le modèle par arbre.....	37
3.2.2.3.1 Surapprentissage	38
3.2.2.3.2 Apprentissage régularisé mais sous-optimum	38
3.2.2.4 Mesure de la performance adaptée au déséquilibre.....	38
3.2.2.4.1 Mesure classique de la performance	38
3.2.2.4.2 Mesures statistiques	39
3.2.2.4.2.1 True Positif Rate ou sensibilité ou encore rappel en français.....	39
3.2.2.4.2.2 Sous-estimation (ou False Negative Rate en anglais)	39
3.2.2.4.2.3 False Positif Rate	39
3.2.2.4.3 F-Mesure.....	40
3.2.2.4.4 Courbe ROC et ROC AUC score.....	40

3.2.2.4.5	Courbe Précision Rappel et score AUC PR.....	41
3.2.2.5	Correction ou non du déséquilibre de classe.....	41
3.2.2.6	Trois niveaux de solutions de correction du déséquilibre	41
3.2.2.6.1	Les solutions de niveau Data.....	41
3.2.2.6.2	Les solutions de niveau algorithmique.....	42
3.2.2.6.3	Les solutions sensibles aux coûts	42
3.2.3	Ajustement du seuil de score de classification	43
3.2.3.1	Méthode de score des probabilités.....	43
3.2.3.1.1	La probabilité d'appartenance à une classe.....	43
3.2.3.1.2	Objectifs des métriques de probabilités	43
3.2.3.1.3	Score Log Loss.....	44
3.2.3.2	Ajustement du seuil de probabilité	44
3.2.3.2.1	Conversion des probabilités à l'étiquette de classe.....	44
3.2.3.2.2	Ajustement du seuil dans le cas d'un déséquilibre de classe	45
3.2.3.2.3	Seuil optimum pour la courbe ROC ou la courbe Précision Rappel ...	46
3.2.3.2.4	Réglage du seuil optimum	46
3.2.3.3	Application de l'ajustement du seuil dans la littérature.....	46
3.2.4	Différentes métriques de l'importance relative des variables	47
3.2.4.1	Interprétabilité	47
3.2.4.2	Explicabilité.....	48
3.2.5	« Trous de connaissances » en classification	49
3.2.6	Aspects « Machine Learning » de la problématique de recherche.....	49
3.3	Prédiction des défaillances avec le projet Signaux Faibles	50
4	Approche empirique de la prédiction de défaillance des entreprises.....	52
4.1	Méthodologie.....	52
4.1.1	Définition de l'échantillon des données des entreprises.....	52
4.1.1.1	Choix du fournisseur des numéros de SIREN	52
4.1.1.2	Mode de sélection des entreprises actives et inactives du secteur industriel.....	52
4.1.1.3	Choix du fournisseur de données comptables et financières	52
4.1.1.4	Sélection des variables explicatives	53
4.1.1.4.1	Variables financières	53
4.1.1.4.2	Variables non financières	56
4.1.1.4.3	Profondeur historique requise	56

4.1.2	Recueil des données.....	57
4.1.2.1	Téléchargement des données de SIREN.....	57
4.1.2.2	Oversampling à des fins de rééquilibrage des classes de défaillances	57
4.1.2.3	Extraction et téléchargement des données comptables et financières ...	57
4.1.2.4	Lecture des fichiers de données.....	57
4.1.3	Visualisation et présentation des données du dataset.....	58
4.1.4	Analyse brute des données	58
4.1.4.1	Qualité des données.....	58
4.1.4.2	Données vides.....	58
4.1.5	Pré-traitement des données	64
4.1.6	Analyse fonctionnelle des données	65
4.1.6.1	Recherche d'une procédure collective.....	65
4.1.6.2	Analyse de la variable ajoutée « horizon de défaillance ».....	67
4.1.7	Constitution du dataset d'entraînement et de test.....	67
4.1.7.1	Sélection opérationnelle des variables cibles et explicatives	67
4.1.7.2	Train_test_split et paramètre stratify	67
4.1.7.3	Shape du dataset de formation et de test	68
4.1.8	Conception du modèle de Machine Learning avec XGBoost.....	68
4.1.8.1	Hyperparamètres du modèle XGBoost	68
4.1.8.2	Choix des autres hyperparamètres importants	68
4.1.8.3	Grid search et recherche des meilleurs paramètres	68
4.1.8.3.1	Validation croisée	68
4.1.8.3.2	Paramètres du GridSearchCV importants	68
4.1.8.4	Recherche des hyperparamètres avec best_estimator_ du GridSearchCV	69
4.1.8.5	Recherche du meilleur seuil de classification	69
4.1.8.5.1	Recherche du seuil optimum pour maximiser la F-Mesure	69
4.1.8.5.2	Recherche du seuil optimum pour minimiser le taux de sous-estimation	69
4.2	Résultats.....	70
4.2.1	Mesure des résultats en termes de performance	70
4.2.1.1	Mesure des résultats « Machine Learning » en termes de performance : Utilisation de la mesure du taux de sous-estimation d'un modèle et méthode d'ajustement d'un seuil de classification en fonction du taux de sous-estimation.	70

4.2.1.1.1	Nombre de défaillances prédites en fonction des Bénéfices et Pertes	70
4.2.1.1.2	Les matrices de confusion résultantes en fonction des différents seuils de classification utilisés	71
4.2.1.1.3	Performances du modèle de prédiction des défaillances.....	71
4.2.1.2	Mesure des résultats « Métier » en termes de performance : Choix de la variable cible	72
4.2.2	Mesure des résultats en termes de volume des prédictions des défaillances	74
4.3	Discussion	75
4.3.1	Discussion des aspects « Machine Learning » : Utilisation de la mesure du taux de sous-estimation d'un modèle et méthode d'ajustement d'un seuil de classification en fonction du taux de sous-estimation	75
4.3.1.1	Différences des mesures statistiques en Médecine et en Machine Learning	75
4.3.1.1.1	Mesures statistiques en Médecine et en Machine Learning	76
4.3.1.1.2	Objectifs des prédictions en Médecine et en Machine Learning.....	77
4.3.1.2	Conclusions sommaires	77
4.3.1.2.1	Résultats de la prédiction avec les différents seuils de classification	77
4.3.1.2.2	Vérification des hypothèses des problématiques de recherche.....	78
4.3.1.3	Pertinence de la recherche.....	78
4.3.1.4	Qualité de la recherche	78
4.3.1.4.1	Validité des résultats	79
4.3.1.4.1.1	Justesse du nombre de Faux Négatifs.....	79
4.3.1.4.1.2	Justesse du nombre de Faux Positifs.....	79
4.3.1.4.1.3	Justesse du seuil de classification calculé pour minimiser le taux de sous-estimation	79
4.3.1.4.2	Pertinence de la méthode	80
4.3.1.4.3	Défauts de la méthode	80
4.3.1.5	Implications pour le domaine de recherche	80
4.3.1.5.1	Comparaison avec les études précédentes.....	80
4.3.1.5.2	Implications pour le domaine.....	80
4.3.1.5.3	Perspectives de recherche	80
4.3.2	Discussion des aspects « Métier »	80
4.3.2.1	Conclusions sommaires	81
4.3.2.1.1	Conclusions intermédiaires concernant les données d'origine	81

4.3.2.1.2 Conclusion finale concernant la prédiction des défaillances d'entreprise	81
4.3.2.1.2.1 Résultats de la prédiction avec la cible « statut juridique »	81
4.3.2.1.2.2 Résultats de la prédiction avec la cible « procédure collective »	84
4.3.2.1.2.3 Vérification des hypothèses des problématiques de recherche	85
4.3.2.2 Pertinence de la recherche.....	85
4.3.2.3 Qualité de la recherche	85
4.3.2.3.1 Validité des résultats	86
4.3.2.3.2 Pertinence de la méthode	86
4.3.2.3.3 Eventuel défaut de la méthode.....	86
4.3.2.4 Implications pour le domaine de recherche	86
4.3.2.4.1 Comparaison avec les études précédentes.....	86
4.3.2.4.2 Implications pour le domaine.....	86
4.3.2.4.3 Perspectives de recherche	86
4.4 Gestion de projet.....	86
4.4.1 Finalités du Proof Of Concept	87
4.4.2 Liste des tâches à effectuer.....	87
4.4.3 Ordonnancement des tâches	87
4.4.4 Durée du projet	87
4.4.5 Jalons & livrables	87
4.4.6 Outils de suivi et de contrôle	88
4.5 Réalisation d'un budget	88
5 Conclusion	89
Bibliographie	90

Table des Illustrations

Figure 1 Nombre de défaillances de PME dans l'industrie depuis 2000	14
Figure 2 Evolution de la production manufacturière depuis 2000	15
Figure 3 Jugement des chefs d'entreprise, dans l'industrie manufacturière, sur leur situation de trésorerie depuis 2008	16
Figure 4 Evolution des algorithmes depuis l'arbre de décision à XGBoost	18
Figure 5 Exemple d'Arbre de décision de comportement sportif : "jouer" en fonction des données météorologiques	18
Figure 6 Le "Bagging" ou "Bootstrap Aggregating"	19
Figure 7 Le "Boosting" ou "Adaptive Boosting"	20
Figure 8 Le "Gradient Boost"	20
Figure 9 Comparatif des performances de XGBoost et des autres algorithmes de ML	21
Figure 10 Surapprentissage de la classe minoritaire	22
Figure 11 Sous-performance des prédictions relatives à une classe minoritaire	22
Figure 12 Mesure de la contribution des facteurs au risque de défaillance	33
Figure 13 exemple d'ajustement du seuil de classification en assurance	47
Figure 14 Le nombre d'entreprises par activité et situation juridique	65
Figure 15 Nombre d'entreprises inactives par situation juridique	66
Figure 16 Nombre d'entreprises en procédures collectives	66
Figure 17 Pertes et Bénéfices des entreprises prédites à risque de défaillance ou non	70
Figure 18 Supériorité du seuil de classification minimisant le taux de sous-estimation	71
Figure 19 Balanced accuracy taux de sous-estimation et ROC AUC du modèle de prédiction des défaillances après différents ajustement du seuil	72
Figure 20 Courbe Précision Rappel du modèle de prédiction des défaillances "Statut juridique"	72
Figure 21 Courbe Précision Rappel et AP du modèle de prédiction "Statut juridique"	72
Figure 22 Courbe ROC et AUC du modèle de prédiction "Statut juridique"	73
Figure 23 Nombre maximum de défaillances prédites en 2022 avec le statut juridique	73
Figure 24 supériorité des mesures des performances des prédictions de défaillances XGBoost seuil / taux de sous-estimation	73
Figure 25 Courbe de Précision Rappel et AP du modèle de prédiction "Procédure Collective"	74
Figure 26 Matrice de confusion en Médecine	76
Figure 27 Matrice de confusion en Machine Learning	76
Figure 28 Courbe Precision Rappel du modèle de prédiction "Statut juridique"	81
Figure 29 Courbe Precision Rappel du modèle de Prédiction "Statut juridique" et AP	82
Figure 30 Courbe ROC et AUC du modèle de prédiction "Statut juridique"	82
Figure 31 Nombre maximum de défaillances prédites en 2022 avec le statut juridique	83
Figure 32 supériorité des mesures des performances des prédictions de défaillances XGBoost seuil / taux de sous-estimation	84
Figure 33 Courbe 1 Précision Rappel du modèle de prédiction des défaillances "procédure collective" abandonné	84
Figure 34 Courbe 2 Précision Rappel du modèle de prédiction des défaillances "procédure collective" abandonné	84
Figure 35 Courbe ROC et AUC du modèle de prédiction 'Procédure collective"	85

Tableau 1 Comparaison du nombre de salariés concernés par une procédure collective ou amiable	30
---	----

Résumé

Ce mémoire propose une application du Machine Learning, plus particulièrement, de la bibliothèque XGBoost, à la prédiction de défaillance des entreprises. Pour cela, une revue de littérature et différents aspects de l'implémentation du modèle sont abordés. La revue de littérature fait apparaître des « trous de connaissances ». Ils portent sur le choix d'une cible alternative de prédiction des défaillances, le choix des variables explicatives, le ratio de déséquilibre des classes toléré, la correction du déséquilibre de classe et l'ajustement d'un seuil de classification pour optimiser la mesure du taux de sous-estimation (False Negative Rate). L'approche empirique décrit la conception du modèle. Ce modèle est basé sur la bibliothèque XGBoost, une bibliothèque de Machine Learning fortement utilisée dans les modèles à succès des compétitions récentes. Les résultats de cette application conduisent à proposer différentes préconisations, à savoir, le choix d'une nouvelle cible de prédiction au travers du statut juridique, l'extension des variables explicatives aux variables non financières (âge de l'entreprise, délais fournisseurs et clients, code NAF, effectif...), la correction du déséquilibre de classe et l'ajustement innovant du seuil de classification en fonction du taux de sous-estimation. Enfin, les résultats de la prédiction de défaillance sont accompagnés d'une mesure de l'importance des variables explicatives. Les récentes techniques de mesure, comme les valeurs de SHAP, offriront de réelles opportunités d'adoption des algorithmes de type « boîtes noires » dans la prédiction des défaillances.

En conclusion, les récentes bibliothèques de Machine Learning, comme XGBoost, peuvent améliorer les résultats de prédiction de défaillances des entreprises, a fortiori, si leur implémentation est couplée avec une vision novatrice du « métier ». Le réexamen de l'objet à prédire et des variables potentiellement explicatives, ainsi qu'une grande connaissance de toutes les possibilités de l'algorithme de Machine Learning utilisé sont nécessaires.

L'explicabilité du modèle reste la condition majeure d'adoption de celui-ci, outre ses performances. Le sujet est particulièrement important pour les modèles de prédiction de type « boîte noire ». La question est : la mesure de l'importance des variables peut-elle convaincre suffisamment les réticents ?

Mots clés :

Défaillance, XGBoost

Abstract

This research essay proposes an application of Machine Learning, more particularly, of the XGBoost library, to the prediction of business failure. To do this, a literature review and various aspects of the implementation of the model are discussed. The literature review revealed various knowledge gaps. They relate to the choice of an alternative failure prediction target, the choice of explanatory features, the tolerated imbalance ratio, the correction of the class imbalance and the adjustment of a classification threshold to optimize the anti-specificity (False Negative Rate). The empirical approach describes the design of the model. This model is based on the XGBoost library, a machine learning library heavily used in successful models of recent competitions. The results of this application lead to propose various and numerous recommendations: firstly, the choice of a new prediction target of legal status, secondly, the extension of the explanatory features to non-financial features (age of the company, supplier and customer deadlines, NAF code, staff, etc.), thirdly, the correction of the class imbalance and fourthly, the innovative of the adjustment of the classification threshold according to the anti-specificity. Eventually, the results of the failure prediction are accompanied by a measure of the importance of the features. Recent measurement techniques, such as SHAP values, will offer real opportunities for the adoption of “black box” algorithms in failure prediction.

To conclude, recent machine learning libraries, such as XGBoost, can improve business failure prediction results. Especially, if their implementation is coupled with an innovative vision of the practice on the job. The re-examination of the object to be predicted, the potentially explanatory features and the great knowledge of all the possibilities of the Machine Learning algorithm used are needed.

The explainability of the model remains the major condition for it to be adopted, together with its performance. The object is particularly important for "black box" type prediction models. The question is: Can the measurement of the importance of features be convincing enough for the reticent?

Key words :

Bankruptcy, XGBoost

Votre admiration m'a donné des ailes...

Loan BONNES,

Fiona BOUNDERIA,

Emmanuelle FABREGUES,

Fabienne FEVRIER,

Florence FEVRIER,

Karine FEVRIER,

Marie-Christine FEVRIER-PERRIN,

Maryse FEVRIER,

Michel PLANCHAIS,

je vous dédie ce mémoire !

*Fais de ta vie un rêve
Et d'un rêve une réalité*

Antoine de Saint Exupéry



*Le Petit Prince, dessin par Antoine de Saint-Exupéry. Domain
public, aucune licence nécessaire.*

Remerciements

Je tiens à exprimer toute ma gratitude à l'égard des personnes qui m'ont aidée à concrétiser ce projet de reconversion professionnelle en Intelligence Artificielle et à bénéficier de cette année d'alternance.

Je remercie ainsi le Ministère des Armées de m'avoir accueillie pendant cette période, si particulière, de la COVID 19 et du télétravail. Mes remerciements s'adressent d'abord à mon tuteur, professeur Michaël KRAJECKI, directeur de projet Intelligence Artificielle et de la Cellule de Coordination de l'Intelligence Artificielle (CCIAD) à l'Agence Innovation Défense (AID), qui m'a recrutée et fait confiance tout au long de cette alternance. Je remercie aussi l'ICETA Emmanuel GARDINETTI, pour sa ténacité, sans qui ce recrutement n'aurait pas été possible. Je remercie également Mattis PAULIN, Architecte, qui dirige le projet et qui m'a guidée dans l'aboutissement de ce Proof Of Concept du Machine Learning en prédiction des défaillances des entreprises. Je remercie Pierre FOULQUIER, chef de bureau, et son responsable David LENOBLE, sous-directeur des PME, pour m'avoir fait confiance lors de la commande de ce projet. Je remercie Yann BOURGEAT, auditeur au service des achats d'armement, pour ses très bons conseils en analyse financière et le temps passé à m'en enseigner les rudiments. Je remercie l'ICA Christine TRICHE, de la CCIAD, pour son empathie, sa philosophie d'entraide et le partage de son carnet d'adresses, qui m'a, de plus, adressée à Jean-Christophe MAURICE, maillon de la chaîne de solidarité menant à Yann BOURGEAT. Je remercie Marie-Véronique SERFATY, comme toute l'équipe de la CCIAD, pour sa bonne humeur égale à son efficacité. Je remercie enfin le personnel de l'AID pour son accueil sincère et tout particulièrement Marie-Christine PILARD, sympathique pilier de la comptabilité de l'AID !

Je tiens également à remercier tous les élèves de la classe M2C de l'IA School, année académique 2020 -2021, pour leur accueil et leur soutien. Je remercie en particulier, mes camarades de projet pour leur disponibilité et leur ténacité pendant les longues heures de visio-conférence et de live-coding. Je citerai en particulier, Haja Nirina ANDRIAMAHEFARIVONY, Lakithan BASKARAN, Fabrice NDOUMBE, Yaya COULIBALY, Jean Marie SADIO, Mohamed KHTEIRA, Mohand Ameziane AIT BEKKA.

Enfin, je remercie Vincent BLOT pour les longues heures d'explications dispensées en Machine Learning, Deep Learning, Computer Vision et TAL, si précieuses pour la réussite de cette année de Master 2 en Intelligence Artificielle.

Globalement, je remercie toutes les personnes qui m'ont aidée de près ou de loin à mettre en œuvre et réussir mon projet de reconversion professionnelle en Intelligence Artificielle à 52 ans. A cet égard, je remercie mes professeurs et particulièrement Lucien AUDIBERT, professeur de mathématiques qui m'a redonné goût aux études et aux mathématiques !

Je remercie également Emmanuelle FABREGUES et Fiona BOUDERIA pour leur amitié et leur soutien dans cette aventure.

Je remercie, aussi, ma famille pour l'amour qu'elle me donne et la confiance qu'elle me porte.

Enfin, je remercie le jury pour l'attention portée à ce mémoire et les différents retours partagés.

1 Introduction

Le Ministère des Armées souhaite suivre et surtout anticiper l'état de santé des entreprises constituant la Base Industrielle et Technologique de Défense (BITD). L'objectif poursuivi est la préservation capacitaire des Armées et pour cela le maintien opérationnel des équipements. Il est donc nécessaire de suivre les fournisseurs dont ils dépendent. La période de pandémie de la COVID 19 a augmenté les risques subis par la BITD et a accentué les besoins de visibilité sur celle-ci. La prédiction des défaillances des entreprises du secteur industriel répond à ce besoin.

Après de fortes inquiétudes sur le climat des affaires dans l'industrie dues au premier confinement, l'heure est à l'optimisme. En juillet 2021, un rapport (Balboni, Boulegue, Mirlicourtois, Passet, & Paturel, 2021) indique, ainsi, que les bases d'une reprise sont présentes : l'offre et la demande ont été préservées, comme les conditions financières nécessaires au bon fonctionnement du marché.

Par ailleurs différentes mesures prises en faveur des entreprises françaises ont eu un fort effet protecteur. Ce qui se traduit par une chute des défaillances d'entreprises depuis 2019 qui se stabilise à un niveau « historiquement bas », en recul sur un an, de 27,6 % pour s'établir à un peu moins de 27.900, selon les dernières statistiques de la Banque de France publiées en juillet 2021 et illustrées en figure ci-dessous (Banque De France Eurosystem, 2021).

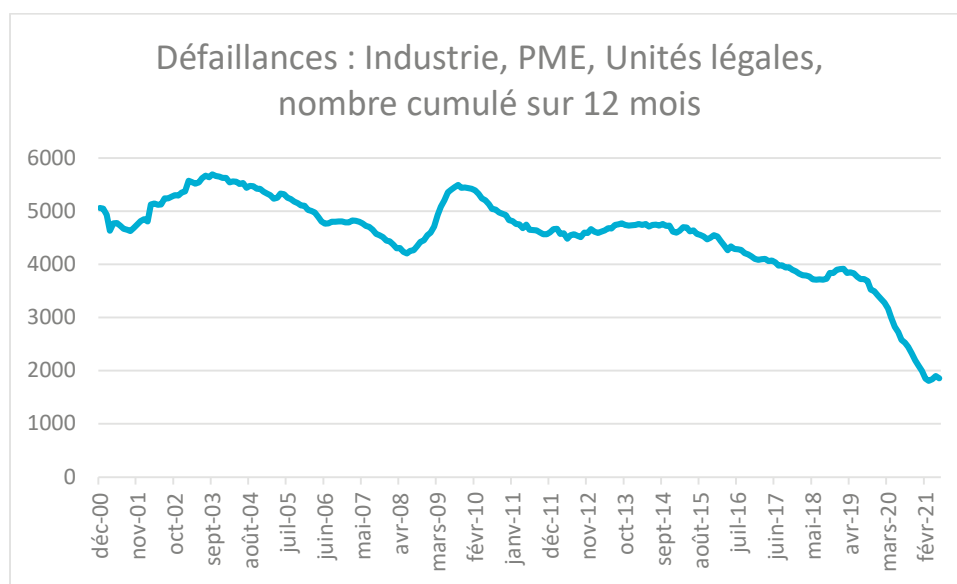


Figure 1 Nombre de défaillances de PME dans l'industrie depuis 2000

Celle-ci explique la baisse des défaillances « par l'impact momentané qu'ont eu les évolutions réglementaires qui ont modifié temporairement les dates de caractérisation et de

déclaration de l'état de cessation de paiements¹ et en second lieu, par l'ensemble des mesures de soutien qui apportent des aides de trésorerie ou permettent aux entreprises de réduire ou retarder le paiement de certaines charges, et donc le risque de faire défaut sur ces paiements (mesures d'activité partielle, prêts garantis par l'État avec remboursements différés, fonds de solidarité, moratoires, etc.). (Banque de France Eurosystem, 2021)

En aout 2021, Xerfi donne un éclairage sectoriel sur l'industrie manufacturière : « Forte du soutien du gouvernement (subventions des investissements, baisse des impôts de production, aide renforcée à l'apprentissage, etc.) et surtout d'une demande domestique et étrangère globalement bien orientée, la production manufacturière continuera de remonter la pente en 2021. Le dynamisme de la demande dans certains marchés, comme le bâtiment et le médical, tirera notamment l'activité vers le haut. Les filières agroalimentaires et les biens d'équipement constitueront également de solides moteurs de croissance, tandis que l'appétence accrue des consommateurs pour le made in France soutiendra l'activité. La situation restera en revanche difficile dans les matériels de transport... ». Xerfi prédit aussi que « dans ces conditions, il faudra attendre 2022 et une nouvelle croissance de 3,4% pour voir l'industrie manufacturière retrouver son niveau d'avant-crise » (Balboni, Boulegue, Mirlicourtois, Passet, & Paturel, 2021)

En juillet 2021, l'INSEE détaille l'évolution de la production manufacturière depuis 2000 par type de production (Laurent, 2021). Elle se révèle contrastée, les catégories « fabrication de matériels de transport » et « autres industries manufacturières » connaissant des difficultés. Ce qu'illustre le graphique suivant :

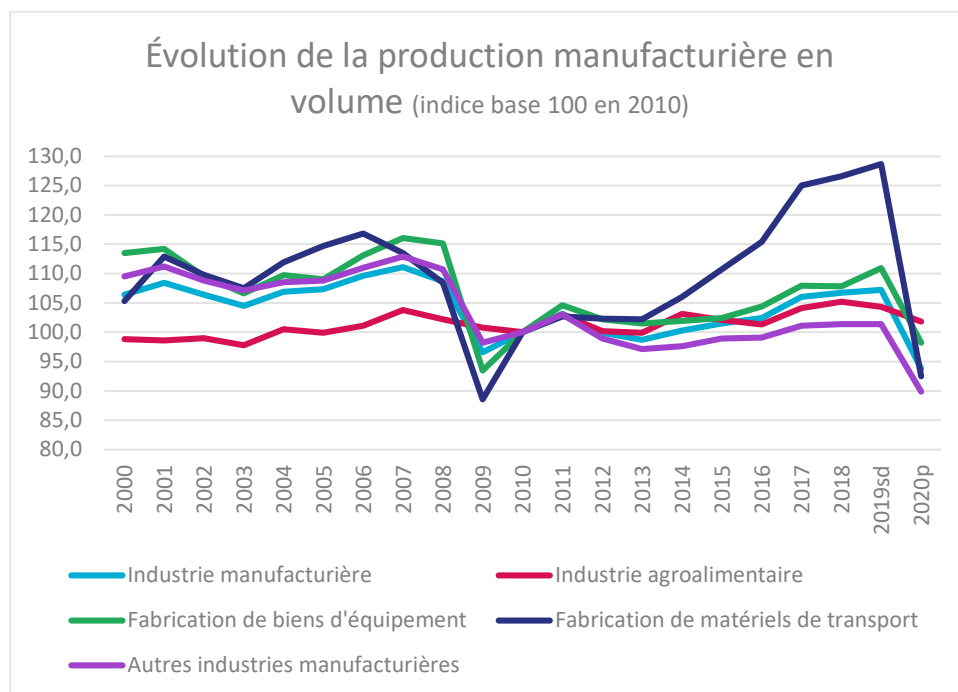


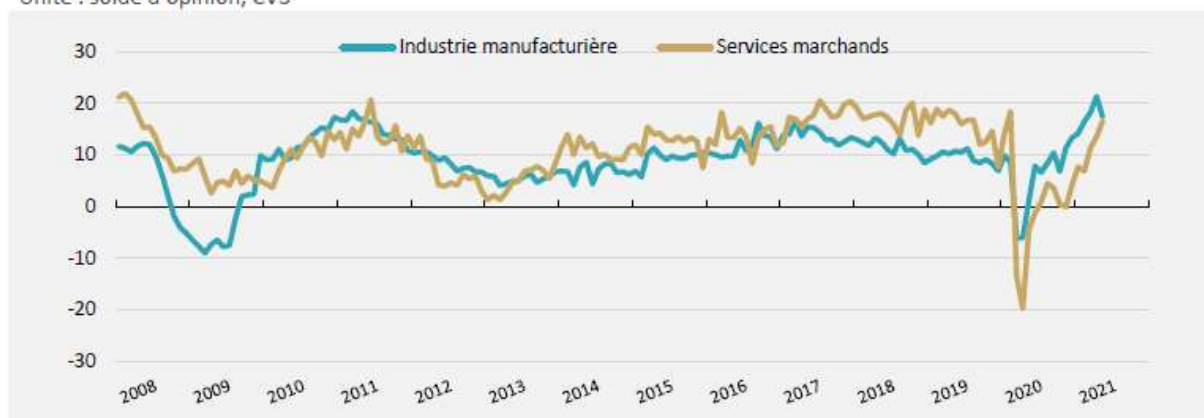
Figure 2 Evolution de la production manufacturière depuis 2000

¹ Appelé aussi dépôt de bilan

Concernant la trésorerie, Xerfi constate que « l'abondance des encaisses de trésorerie, non mobilisées à ce stade, réduit le stress du paiement des charges différées de la crise (taxes, loyers différés, remboursement du PGE²), et par là même les risques de défaillances. Si l'on ajoute à cela la possibilité de prolonger de 5 à 6 ans le dispositif du PGE, à très faible coût de crédit, nombre d'entreprises choisiront sans doute de conserver leur aisance en *cash*, soit comme matelas de sécurité, soit pour financer de nouveaux investissements ou des redéploiements. De fait, le PGE est en passe de se transformer en « assurance-risque » et en aiguillon d'audace entrepreneuriale, en appui de la reprise. » (Balboni, Boulegue, Mirlicourtois, Passet, & Paturel, 2021). Les auteurs ajoutent que « la perception des entreprises est en ligne avec leur situation bancaire. L'action des aides publiques conjuguée à l'amélioration de l'activité économique a fait mieux que préserver la trésorerie. Les dépôts bancaires des entreprises surpassent aujourd'hui largement leurs niveaux d'avant-crise. Autrement dit, beaucoup ont constitué un matelas de sécurité. Ce constat amène à tempérer le risque de grande vague de défaillances à court terme et augure d'une capacité de résistance renforcée dans le cadre de l'étirement de la crise sanitaire. L'ampleur du soutien public a mis en échec les pronostics les plus noirs de défaillances en chaîne des entreprises fin 2020 ou début 2021 » (Balboni, Boulegue, Mirlicourtois, Passet, & Paturel, 2021). Ce qu'illustre la figure suivante :

Jugement sur la situation de trésorerie

Unité : solde d'opinion, CVS



Source : Banque de France, dernière donnée disponible 05/2021

Figure 3 Jugement des chefs d'entreprise, dans l'industrie manufacturière, sur leur situation de trésorerie depuis 2008

Mes travaux commencent, au chapitre 2, par décrire le contexte du projet de prédiction des défaillances. Ils se poursuivent par une revue de littérature dans le domaine de la prédiction des défaillances et de l'état de l'art en classification, ainsi qu'une étude des résultats et de

² Une entreprise dont la trésorerie est impactée par l'épidémie de coronavirus - Covid-19 peut demander un prêt garanti par l'État (PGE), quelle que soit sa taille et son statut. Cette aide s'applique jusqu'au 31 décembre 2021.

l'approche du projet « Signaux Faibles », un outil de prédiction des entreprises en risque de défaillance (République Française Data.gouv.fr, 2021).

La revue de littérature débute par un rappel sur la notion de défaillance. En 2019, l'INSEE rappelle sa définition de la défaillance d'entreprise à savoir : « Une unité légale est en situation de défaillance ou de dépôt de bilan à partir du moment où une procédure de redressement judiciaire est ouverte à son encontre. Cette procédure intervient lorsqu'une unité légale est en état de cessation de paiement, c'est-à-dire qu'elle n'est plus en mesure de faire face à son passif exigible avec son actif disponible. ». (INSEE, 2019)

Néanmoins, l'INSEE, insiste sur la différence entre les notions de défaillance et de cessation d'activité (INSEE, 2019), cette dernière notion étant peu exploitée dans les travaux de recherche sur le sujet.

La notion de défaillance a été étudiée dans plusieurs publications dans différents pays. Les raisons pour lesquelles ces études ont été conduites sont essentiellement économiques et financières. C'est donc sous cet angle, et ses conséquences, que la prédiction a été conçue. Historiquement, celle-ci a été réalisée avec des outils mathématiques statistiques principalement, ce qui a conduit à la conception de nombreux calculs de ratios financiers pour classer les performances des entreprises et leur état de santé et, in fine, prédire leur survie. En 2004, un état des lieux concernant la prévision de la faillite a été publié (Refait-Alexandre, 2004). Il retrace l'histoire de la prévision de défaillance et les différentes technologies appliquées au cours du temps, ainsi que les quatre grandes étapes successives de la prédiction : construction de l'échantillon, sélection a priori des variables explicatives, choix du mode de classification et estimation de la qualité de la prévision. Cette historique remonte aux premiers travaux d'analyse unidimensionnelle (Beaver, 1966) et surtout l'analyse discriminante multidimensionnelle (Altman, 1968) jusqu'aux premiers travaux de prédiction de défaillance en intelligence artificielle : comparaison avec les réseaux de neurones (Altman, Marco, & Varetto, 1994), (Bardos & Zhu, 1997), et algorithmes génétiques (GA) (Varetto, 1998). Cet état des lieux précise, notamment, les notions de classes de défaillances et de facteurs explicatifs utilisés.

Plus globalement, cette revue de littérature présente les facteurs explicatifs usuellement exploités pour prédire la défaillance ainsi que les facteurs explicatifs potentiels peu considérés jusqu'à présent. En effet, certains facteurs explicatifs ont été confirmés par une étude en 2010 notamment (Levratto, Carré, Zouikri, & Tessier, 2011), sans être exploités lors des travaux ultérieurs de prédiction de défaillance. Cette désaffection peut trouver son origine dans l'absence de données concernant ces facteurs.

Par ailleurs, le Machine Learning s'est révélé être un puissant outil dans le domaine de la data science, mais pas seulement. C'est le cas de l'algorithme XGBoost, dans le domaine de la prédiction des défaillances. Cet algorithme est basé sur les arbres de décision. Ces derniers ont fortement évolué pour accroître leurs performances. Les conceptions des algorithmes de Decision Tree, de Bagging, de Random Forest, de Boosting, de Gradient

Boosting et de XGBoost (Extreme Gradient Boosting) ont constitué les étapes intermédiaires de la recherche du meilleur algorithme (Morde, 2019).

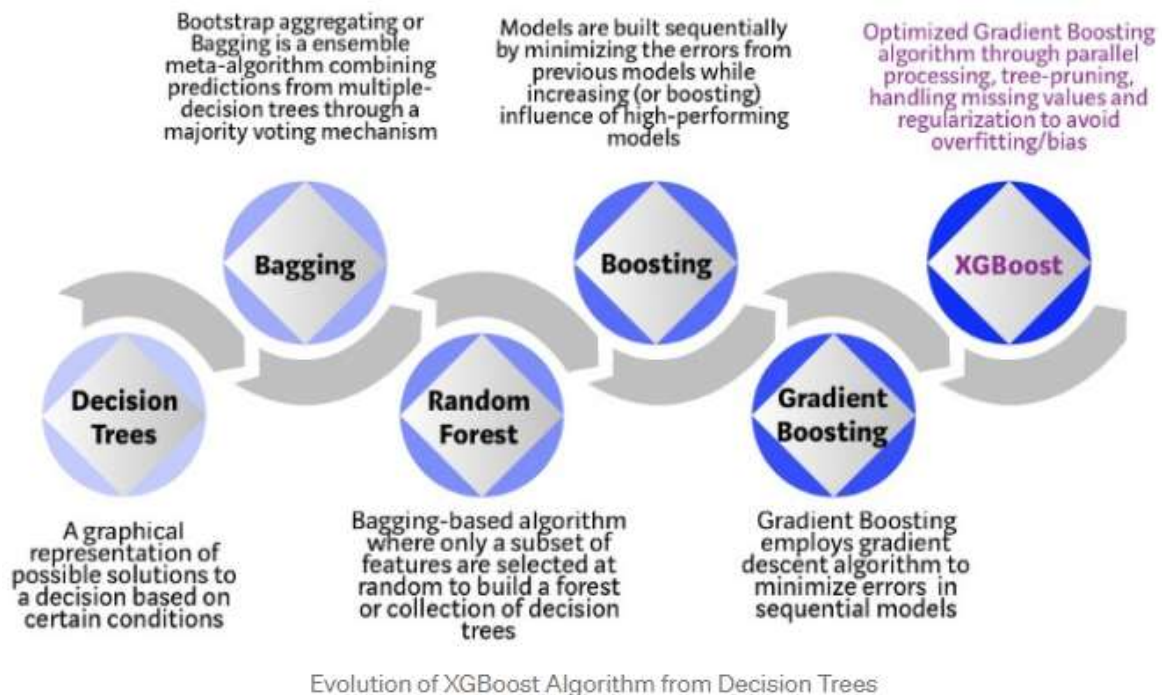


Figure 4 Evolution des algorithmes depuis l'arbre de décision à XGBoost

Pour rappel, l'apprentissage par arbre de décision (Decision Tree) utilise un arbre de décision comme modèle prédictif. « C'est une technique d'apprentissage supervisé : on utilise un ensemble de données pour lesquelles on connaît la valeur de la variable-cible afin de construire l'arbre (données dites étiquetées), puis on extrapole les résultats à l'ensemble des données de test. » (Wikipédia L'encyclopédie libre, 2021).

La figure suivante illustre le cas de la prédiction de comportement des sportifs : « jouer » en fonction des données météorologiques (Wikipédia L'encyclopédie libre, 2020).

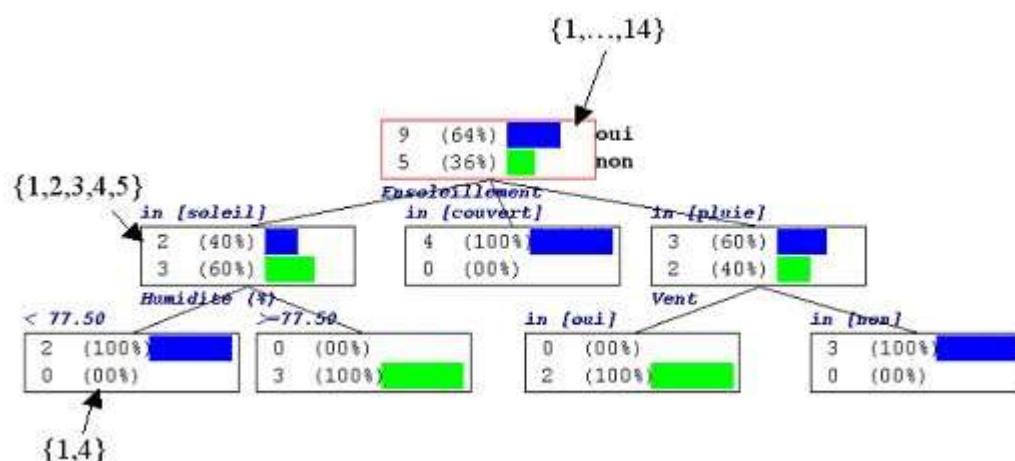


Figure 5 Exemple d'Arbre de décision de comportement sportif : "jouer" en fonction des données météorologiques

La variable de comportement sportif « jouer » figure à chaque nœud de l'arbre. Dans notre cas, sur les 14 observations recensées et dénombrées à la racine, 9 observations

appartiennent à la classe « oui », en bleu, indiquant qu'une partie a eu lieu, et 5 observations appartiennent à la classe « non », en vert, indiquant qu'aucune partie n'a eu lieu. Chaque valeur des variables météorologiques est ensuite exploitée pour créer des sous-ensembles homogènes. Ainsi, la variable « ensoleillement » est utilisée pour séparer l'ensemble des données (dataset en anglais) en 3 nœuds suivant les valeurs de la variable. Dans notre exemple, le nœud fils le plus à gauche (« ensoleillement » = soleil) comporte 5 observations au total, dont 2 appartenant à la classe « oui » et 3 à la classe « non ». Toutes les variables météorologiques sont, ainsi de suite, exploitées pour créer des nœuds de plus en plus homogènes vis-à-vis de la classe à prédire « jouer ». Dans l'idéal, les feuilles portent un sous-ensemble homogène : dans notre exemple 100% de la classe « oui » ou 100% de la classe « non ». L'examen des variables est propre aux différents algorithmes de classification.

Une première évolution de l'arbre de décision, le « Bagging » ou le « Bootstrap Aggregating », a consisté à assembler plusieurs arbres de décision. Le principe est d'assembler, en parallèle, plusieurs arbres de décision dits faibles (weak learner) pour constituer un modèle plus performant (strong learner) (Chen, 2019). Une procédure, dite « Bootstrap », permet de séparer les données d'entraînement en différents sous-échantillons aléatoires, que différentes itérations du modèle utilisent pour s'entraîner.

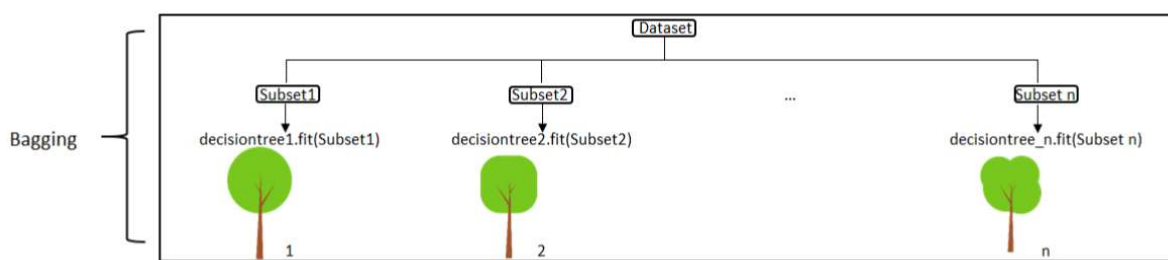


Figure 6 Le "Bagging" ou "Bootstrap Aggregating"

Les résultats de chaque arbre de décision faible sont combinés pour obtenir un super résultat final, par un vote de la classe à prédire par chaque arbre. Le résultat final est la classe comportant le vote maximum.

L'algorithme « Random Forest » est un type d'algorithme « Bagging » particulier où le choix des variables à tester à chaque nœud est fait pour maximiser le gain d'informations, c'est-à-dire l'homogénéité des sous-ensembles déduits, que l'on obtiendrait si l'on choisissait cette variable (Elbaz, 2020).

Une seconde évolution, le « Boosting », ou « Adaptive Boosting », a consisté à séquencer les arbres de décision faibles, chacun pour corriger le résultat du précédent. (Chen, 2019)

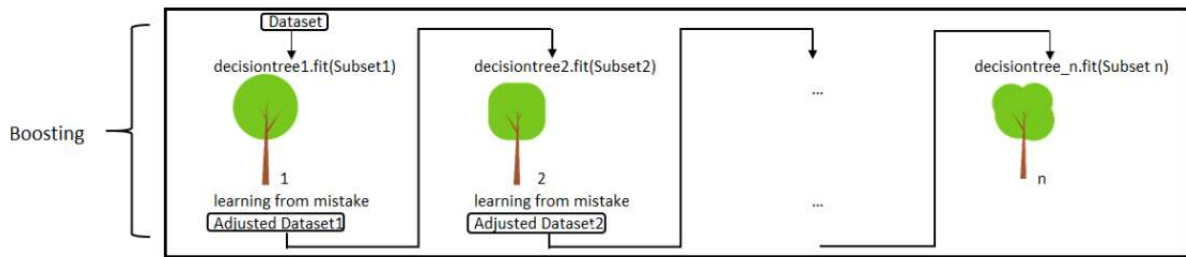


Figure 7 Le "Boosting" ou "Adaptive Boosting"

Les différents arbres de décision faibles sont pondérés de telle sorte qu'à chaque prédiction, les arbres ayant mal prédit auront un poids plus faible que ceux dont la prédiction est correcte.

Une troisième évolution, « Gradient Boosting », a consisté à « apprendre » les erreurs résiduelles et à les minimiser.

Le premier arbre de décision faible effectue la moyenne des observations et sert de base au reste des arbres de décision faibles. Le second arbre de décision prend en compte l'écart entre la moyenne et la réalité, appelé « erreur résiduelle », et cherche à la prédire et à la réduire. Le troisième arbre faible prend en compte l'erreur résiduelle du deuxième et ainsi de suite (T, 2020).

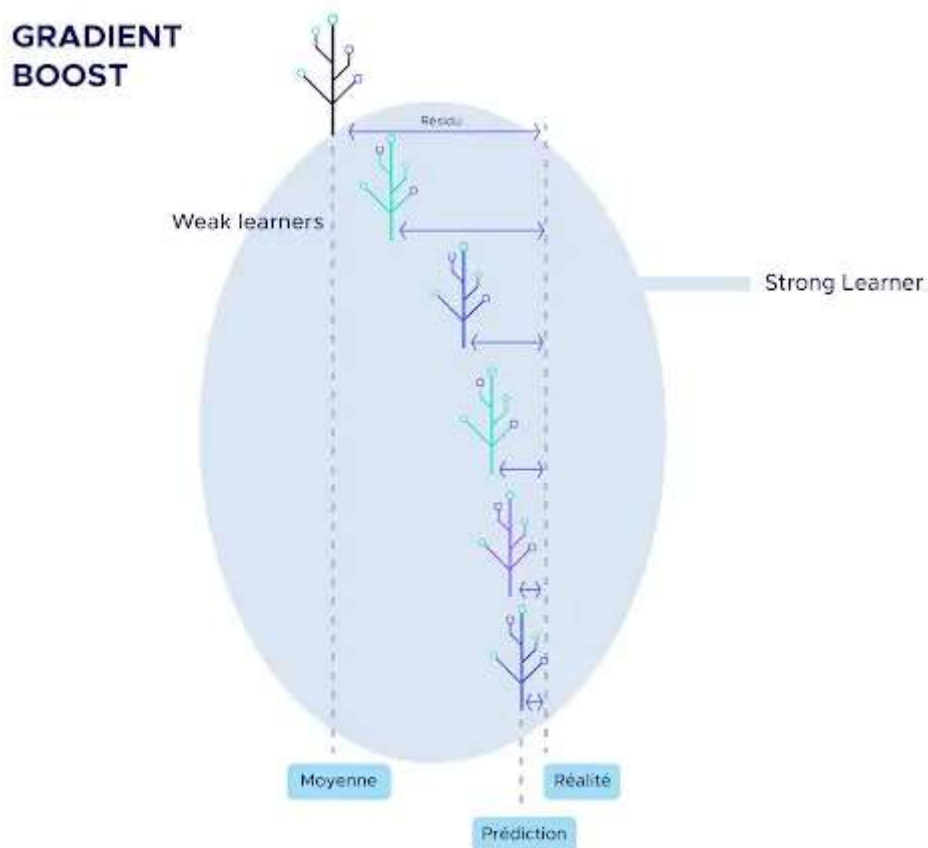


Figure 8 Le "Gradient Boost"

Enfin, l'algorithme XGBoost, est un algorithme de type « Gradient Boost ». Sa particularité est d'élaguer les arbres de décision « trop » faibles pour les rendre bons ou ne conserver que ces derniers. C'est la technique du « pruning ». Par ailleurs, cet algorithme est optimisé pour raccourcir les temps de calcul. De plus, cet algorithme offre de nombreux hyperparamètres, le rendant particulièrement adaptatif. C'est pourquoi XGBoost est souvent choisi pour concevoir les modèles de Machine Learning gagnant les compétitions.

En 2017, une étude a permis de comparer les résultats de différents modèles de Machine Learning de prédiction des défaillances, basés sur des algorithmes de SVM, bagging, boosting, et Random Forest (Barboza, Kimura, & Altman, 2017).

En 2019, un autre article a donné le comparatif des performances de plusieurs algorithmes de classification et a confirmé la suprématie de la bibliothèque XGBoost (Morde, 2019).

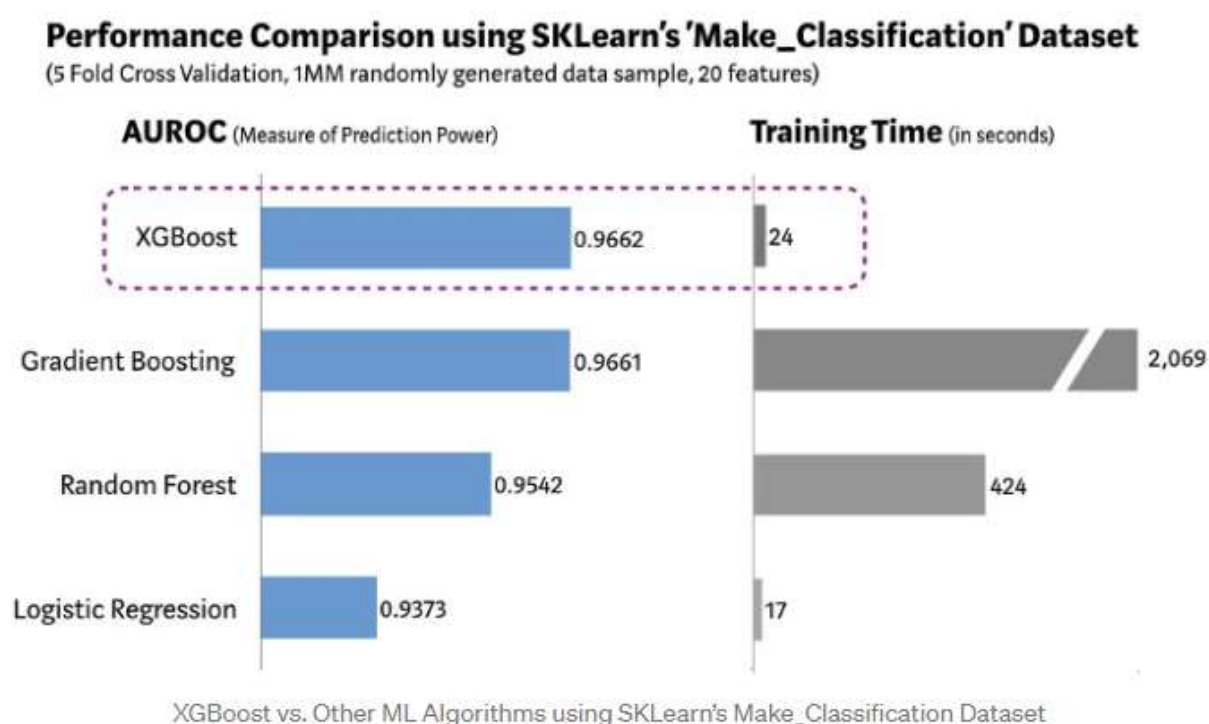


Figure 9 Comparatif des performances de XGBoost et des autres algorithmes de ML

Plus récemment, en 2021, une étude a été publiée sur les différentes approches de Machine Learning et de statistiques pour la prédiction des faillites bancaires aux USA et en Europe (Durand & Le Quang, 2021). Cette étude compare les performances des algorithmes de type régression linéaire, forêt aléatoire et réseau de neurones, pour les deux continents, et conclut sur la supériorité du réseau de neurones pour les données américaines.

La revue de littérature se poursuit par un zoom sur les deux aspects suivants : le traitement du déséquilibre des classes et le seuillage du score de la classe de défaillance prédite.

En 2021, un article de blog présente les différents aspects et les conséquences du déséquilibre des données, ainsi que les méthodes de traitement (Tremblay, 2021). Cet article

illustre bien le risque de surapprentissage des algorithmes sur la base d'entraînement et donc de mauvaise prédiction sur la base de test.

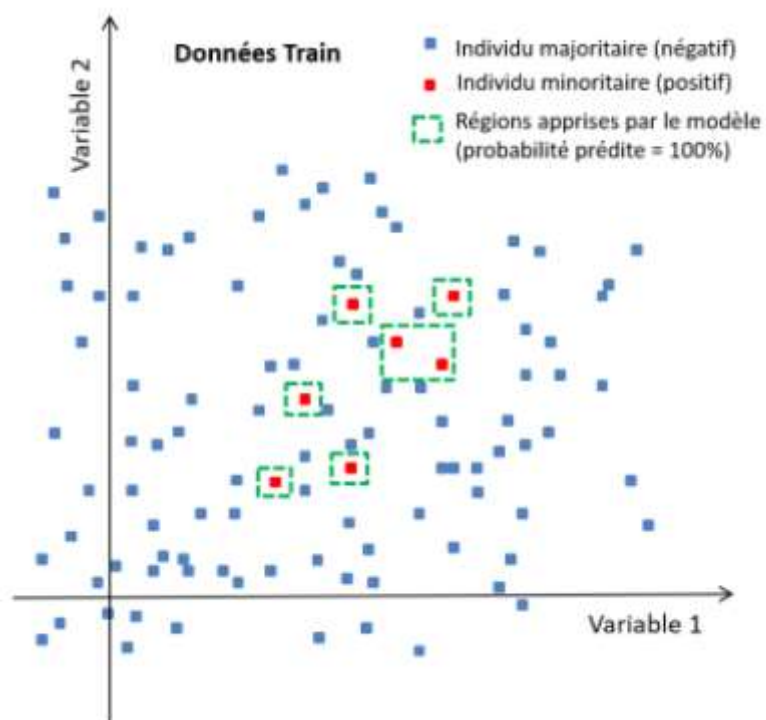


Figure 10 Surapprentissage de la classe minoritaire

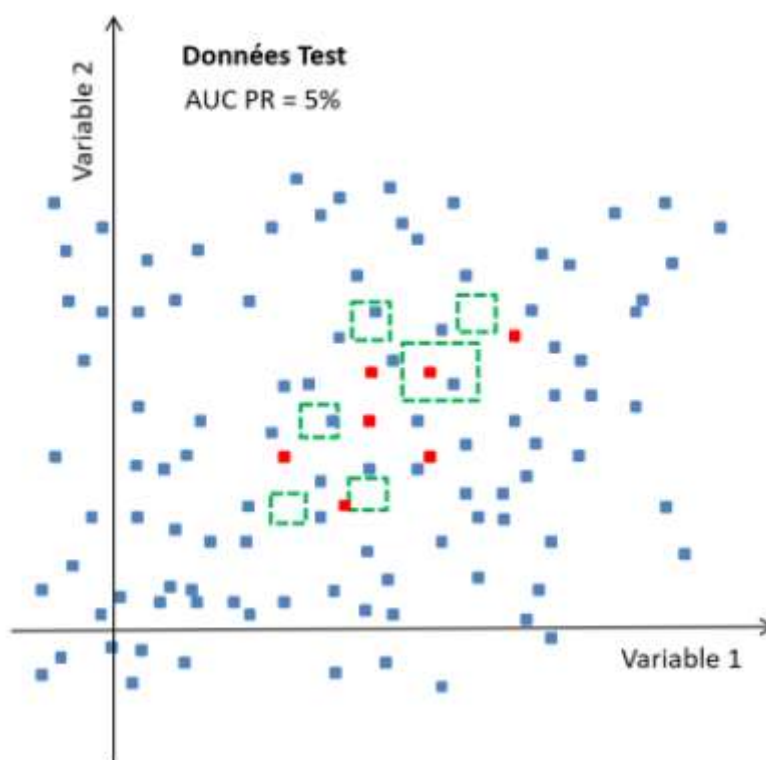


Figure 11 Sous-performance des prédictions relatives à une classe minoritaire

L'article met en garde contre la solution « naïve » de régularisation qui réduit la variance sans réellement améliorer les résultats. Par ailleurs, il conseille d'adapter les métriques d'évaluation des résultats, pour tenir compte des effets des classes déséquilibrées. Enfin, trois familles de méthodes correctrices sont indiquées « pour apprendre à partir de données déséquilibrées : les méthodes de resampling qui modifient les données brutes d'apprentissage, les méthodes d'apprentissage spécifique avec rééchantillonnage interne ou pondération, et les méthodes d'apprentissage sensible aux coûts (cost-sensitive) » (Tremblay, 2021).

Ce traitement du déséquilibre de classe a, notamment, été étudié dans le domaine de la détection de fraudes. En effet, la fraude est sous-représentée par rapport au cas de non-fraude, malgré un coût énorme. Une première thèse en 2019 apporte quatre approches pour traiter ce déséquilibre (Metzler, 2019). Une thèse professionnelle aborde aussi ce sujet en 2021. Cette thèse émet plusieurs hypothèses concernant l'efficacité des méthodes « cost-sensitive » (méthode d'ajustement du coût d'un mauvais classement), la performance des algorithmes de Boosting et l'efficacité des algorithmes ensemblistes par rapport aux algorithmes non-ensemblistes (Alfocea, 2021)

En ce qui concerne le seuillage du score de la classe prédite, une thèse en 2019 présente un comparatif des performances en fonction du choix du seuil de classification (Ly, 2019). En effet, par défaut le seuil de classification est fixé à 0,5. Il peut être opportun de le modifier suivant le résultat recherché.

La revue de littérature se termine par l'état de l'art de l'interprétabilité des modèles de Machine Learning donné dans un billet de blog fin 2020 (aquila data enabler, 2020). En effet, les résultats de certains modèles complexes performants, encore appelés « black box », doivent pouvoir être accompagnés d'informations sur les paramètres qui ont influencé chaque valeur prédite. Il s'avère utile (ou nécessaire) d'être en mesure d'expliquer la prédiction pour les raisons suivantes :

- Pour le Data Scientist :
 - Pour déterminer quand et pourquoi le modèle se trompe,
 - Pour valider la cohérence du modèle,
 - Pour donner confiance au métier vis-à-vis du modèle conçu,
- Pour le métier :
 - Pour pouvoir justifier les décisions prises (auprès d'un client ou d'une direction),
 - Pour gagner en transparence,
 - Pour être conforme avec les réglementations en vigueur (RGPD, aspect légal, audit, etc...).

Le billet explore les techniques de mesure de l'importance des variables et conclut sur une mise en garde de l'utilisation des fonctions de mesure « de base ». Il préconise les solutions alternatives de la « permutation feature importance » (différence des performances lors des permutations de variables) ou des valeurs SHAP (SHapley Additive exPlanation) (contributions de chaque variable à la prédiction).

Le présent mémoire se poursuit par la présentation du projet de prédiction des défaillances avec le produit « Signaux Faibles » (République Française, 2021).

Ce produit, en accélération, a pour objectif de mieux cibler les interventions en remédiation de l'Etat vers les entreprises en difficulté. L'équipe est portée par le Ministère de l'Économie, des Finances et de la Relance (DGE) en partenariat avec le Ministère du Travail, de l'Emploi et de l'Insertion (DGEFP), la Banque de France, l'Agence centrale des organismes de sécurité sociale (ACOSS) et la Direction interministérielle du numérique (DINUM). Ce projet est une réalisation de l'incubateur du Ministère de l'Économie, des Finances et de la Relance (MEFR). Le projet résout le problème de la découverte tardive des difficultés d'entreprises par les CRP (Commissaires aux Restructurations et à la Prévention des difficultés des entreprises). Sa solution est de valoriser la richesse des données administratives des partenaires du projet pour produire un outil d'analyse prédictive des difficultés des entreprises.

Plus précisément, le produit « Signaux Faibles » prédit un risque d'entrée en procédure collective des entreprises actuelles, à partir de leurs données passées et des trajectoires des entreprises ayant connu une défaillance. Il s'appuie sur un modèle récent de Machine Learning, pour fournir une prédiction de défaillance à 18 mois pour les entreprises qui disposent d'établissements de plus de 10 salariés. Une liste des entreprises détectées en difficulté est transmise aux différentes administrations partenaires pour activation des leviers d'accompagnement propres à chacune.

Le code source est libre et fait l'objet d'un dépôt sur GitHub (GitHub / Signaux Faibles, 2021). Ce dépôt comporte, notamment, la liste des données utilisées par le produit (Camilleri, Joli, Viers, Coufour, & Ninucci, 2021). Cette liste a inspiré la liste des variables exploitées dans l'approche empirique de prédiction des défaillances avec la bibliothèque XGBoost.

Une approche empirique a été menée d'une part pour répondre à la demande de visibilité du Ministère des Armées, et d'autre part pour valider, par un Proof Of Concept (POC), la méthodologie de la prédiction des défaillances des entreprises dans le nouveau contexte de la pandémie. Le 4^{ème} chapitre du mémoire y est consacré.

La définition de l'échantillon et méthodologie du recueil d'informations débutent l'approche empirique de ce mémoire. Elles décrivent la constitution de l'échantillon des entreprises et la profondeur de l'historique des données choisies. L'échantillon de données du POC a été constitué indépendamment du contour de la BITD. Néanmoins, l'échantillon a été conçu pour être relativement représentatif du monde industriel français, et être au plus proche de la BITD. Par ailleurs, pour améliorer l'apprentissage du modèle de prédiction, des entreprises ayant connu des défaillances ont été insérées dans l'échantillon. Pour les mêmes raisons, une profondeur historique relative de 4 ans a été choisie, que l'entreprise ait cessé ou pas son activité.

Le recueil des données du POC a été réalisé, à partir de la base open data Siren® (INSEE, 2021) et d'un fournisseur externe privé de données comptables et financières.

Les données recueillies sont ensuite visualisées. Cette phase a répondu au besoin de vérification de la qualité des données, en particulier du taux de présence / absence des données. En effet, les fournisseurs de données comptables et financières sont généralement tributaires de la publication des comptes financiers, hors exigence de confidentialité, car depuis la loi MACRON de 2014, Décret n° 2014-1189 du 15 octobre 2014 relatif à l'allègement des obligations de publicité des comptes annuels des micro-entreprises (Légifrance le service public de la diffusion du droit de la République Française, 2014), modifiée plusieurs fois depuis, certaines entreprises ont la possibilité de rendre confidentiel leur dépôt de comptes auprès du greffe du tribunal de commerce afin que ces comptes ne soient pas rendus publics. Depuis la loi PACTE du 23 mai 2019 (CCI Paris Ile De France, 2019), l'option de confidentialité :

- Des comptes sociaux est réservée aux micro-entreprises remplissant au moins 2 des critères suivants :
 - Total de bilan de moins de 350 000 euros,
 - Chiffre d'affaires net de moins de 700 000 euros,
 - Effectifs inférieurs à 10 salariés.
- Du compte de résultat seulement, le bilan restant public, s'applique aux petites entreprises dont les comptes sociaux sont publiés depuis le 7 août 2016 et ne dépassant pas au moins 2 des critères suivants :
 - Total du bilan de 4 millions d'euros,
 - Chiffre d'affaires net de 8 millions d'euros,
 - 50 salariés.

Dans ce cas, seules les administrations, les autorités judiciaires ou la Banque de France peuvent y avoir accès.

Par ailleurs, cette visualisation des données comptables et financières fait apparaître un déficit d'information des entrées en procédure collective par rapport aux cessations d'activité.

Concernant la mise au point du modèle de prédiction, une validation croisée a été utilisée pour la recherche des meilleurs hyperparamètres de l'algorithme XGBoost en raison de la petite taille du dataset disponible ; les données de 2 348 entreprises, sur les 3 000 requises, ont été finalement recueillies en phase de POC.

La conception du modèle a d'abord été centrée sur le choix de la cible à prédire et sur le choix des variables potentiellement explicatives.

Ensuite, le déséquilibre de la classe de défaillance prédite a été traité. En effet, malgré l'insertion de données relatives à des entreprises supplémentaires ayant cessé leur activité dans le passé, les données de défaillance sont largement minoritaires. C'est pourquoi, une pondération des classes de défaillance a été effectuée. Par ailleurs, la définition de trois seuils différents de scores de classification de défaillance permet d'ajuster les performances du modèle de prédiction et la liste des entreprises à risque de défaillance établie.

La partie empirique du mémoire se poursuit par la restitution des résultats. Ces résultats font, ensuite l'objet d'une discussion poussée et détaillée pour chacune des hypothèses émises.

Le mémoire conclut, enfin sur les réponses apportées aux problématiques de recherche étudiées.

2 Contexte du projet de prédiction des défaillances

Un des objectifs de la Direction Générale de l'Armement (DGA) est la préservation capacitaire des Armées et pour cela le maintien en condition opérationnelle des équipements. Ceci se traduit par une demande de visibilité sur l'état de santé futur des fournisseurs des équipements des Armées. Cette demande s'est trouvée renforcée dans la période de crise sanitaire actuelle.

Ce projet répond à la demande de visibilité par la prédiction des défaillances des entreprises de la BITD, au moyen d'un modèle de Machine Learning basé sur une bibliothèque logicielle open source XGBoost puissante et récente.

Cette prédiction fournit une liste d'entreprises signalées à risque de défaillances dans les prochains mois (l'horizon de défaillance constaté dans le passé est de 18 mois en moyenne).

Le POC, constituant l'approche empirique de ce mémoire, a pour objectif de valider le modèle de prédiction de défaillance sur un ensemble de données de test et de vérifier la qualité de la réponse au besoin de la DGA. Le passage à l'échelle sur les données des entreprises de la BITD sera mené ultérieurement, en fonction des résultats du POC.

Il est à noter que d'autres organisations (Banque de France, INSEE, Xerfi, ...) (cf. Introduction) produisent ce type d'informations au niveau global des entreprises françaises ou parfois plus précisément au niveau de certains secteurs industriels, ce qui ne répond pas complètement au besoin de la DGA.

3 Revue de littérature

Avant de mener le POC de prédiction des défaillances des entreprises, un état de l'art et une revue de littérature des domaines concernés, à savoir la défaillance d'entreprise et les algorithmes de classification en Machine Learning, sont opportuns.

3.1 Cadre théorique de la défaillance d'entreprise

3.1.1 Notion de défaillance d'entreprise

3.1.1.1 Différents événements de la vie d'une entreprise

Pour mémoire, une entreprise est schématiquement en procédure de redressement judiciaire lorsqu'elle est en état de cessation de paiement, caractérisé par l'impossibilité de faire face au passif exigible (les dettes) avec l'actif disponible (INSEE, 2019). Néanmoins, il convient de détailler cette notion de défaillance d'entreprise en termes de procédures préventives d'une part et de procédures collectives d'autre part.

La vie d'une entreprise n'est pas un long fleuve tranquille : conjoncture économique difficile, mauvaise étude de marché, business plan mal cadré, impayés de la part de certains des clients et mauvaise gestion peuvent déboucher sur des difficultés financières. Lorsqu'une entreprise rencontre ces difficultés ou prévoit d'en rencontrer à court terme, différentes procédures, ayant des objectifs propres, peuvent être mises en place.

Le traitement des difficultés a évolué dans le temps au profit des entreprises en difficulté : le législateur a mis en place un arsenal juridique de moins en moins « sévère et éliminateur », et de moins en moins marqué par l'unilatéralisme de l'autorité judiciaire, d'après une thèse étudiant la négociation en droit des entreprises en difficulté (Koehl, 2019).

En amont, des tentatives de prévention ou de résolution amiable des difficultés peuvent être opérées en lançant, à la demande du chef d'entreprise, une des procédures préventives amiables : le mandat ad hoc³ ou la procédure de conciliation⁴, ou bien la première procédure collective : la procédure de sauvegarde judiciaire. Les procédures préventives amiables et de sauvegarde judiciaire ont pour but d'aider le débiteur à faire face à ses créanciers tout en maintenant les emplois, le tissu économique ou en réorganisant l'entreprise. La négociation débiteur – créanciers est au cœur de ces procédures. En effet, l'entreprise sort de l'état de cessation de paiement, lorsqu'elle bénéficie de réserves de crédits supplémentaires ou lorsqu'elle obtient un délai de paiement de la part de ses créanciers, car cela a pour conséquence de lui permettre de faire face, à nouveau, au passif exigible grâce à son actif disponible (Ministère de l'économie des finances et de la relance Bercy Entreprises Infos, 2019).

³ Un administrateur judiciaire, justifiant d'une grande expérience dans le domaine des entreprises en difficulté est nommé en tant que mandataire

⁴ Un conciliateur est nommé par le président du tribunal.

Il est à noter que parmi ces procédures préventives, seule la procédure de sauvegarde judiciaire est publique, les deux autres étant couvertes par le sceau de la confidentialité. Par ailleurs, toutes ces procédures s'appliquent suffisamment tôt dans la rencontre des difficultés, à savoir avant l'état de cessation de paiement. Néanmoins la procédure de conciliation comporte une particularité : son ouverture est autorisée jusqu'au 45^{ème} jour d'état de cessation de paiements.

Ensuite, lorsque les difficultés s'avèrent trop importantes ou lorsque la négociation est en échec, les créanciers peuvent demander l'ouverture d'une procédure collective plus contraignante, à savoir la procédure de redressement judiciaire. Elle commence par une période d'observation, puis, se termine par un plan de redressement si l'entreprise est viable. Dans le cas contraire, elle débouche sur une liquidation judiciaire.

La liquidation judiciaire intervient à la suite d'une procédure de sauvegarde ou de redressement judiciaire ou lorsque ces dernières ne sont plus envisageables.

La clôture de la liquidation judiciaire place l'entreprise en cessation d'activité. Elle peut aussi faire l'objet d'une reprise globale ou partielle (Ministère de l'économie, des finances et de la relance, Bercy Infos, 2019).

Pour conclure, les procédures préventives et amiables étant confidentielles, hormis la procédure collective de sauvegarde de justice, peu d'informations émanent des entreprises rencontrant les prémices de leurs difficultés financières. Ces informations sont disponibles dans les services juridiques ou contentieux seulement.

3.1.1.2 Hypothèse de l'impact des procédures amiables sur la production industrielle

Par ailleurs, ces procédures préventives et amiables concernent beaucoup de salariés malgré leur nombre moindre par rapport aux procédures collectives, comme le montrent les données de source Deloitte Altares de 2018 dans une thèse portant sur la négociation en droit des entreprises en difficulté (Koehl, 2019).

Tableau 4 - Salariés concernés par les procédures collectives (2012-2017)

	2012	2013	2014	2015	2016	2017
Salariés	268 452	272 714	245 589	234 453	193 649	171 667

Source : Deloitte **Allures**, avril 2018.

Tableau 5 - Salariés concernés par les procédures amiables -mandat ad hoc et conciliation- (2012-2016)²⁶⁵⁸

	2012	2013	2014	2015	2016
Salariés	495 901	551 570	561 452	582 435	612 001

Source : Deloitte **Allures**, avril 2018.

Tableau 1 Comparaison du nombre de salariés concernés par une procédure collective ou amiable

On peut raisonnablement supposer qu'il en est de même sur le volume de la production en jeu. Ainsi le volume de la production industrielle menacée lors des procédures amiables serait plus grand que lors des procédures collectives. Cela reste à confirmer faute d'études statistiques publiées sur ce sujet.

3.1.1.3 Cessation d'activité d'une entreprise

Il est opportun de clarifier la distinction entre les notions de défaillance et de cessation d'activité. La notion de cessation correspond à l'arrêt total de l'activité économique d'une entreprise, identifiée par son numéro Siren®. Toutes les défaillances ne débouchent pas sur des cessations. Par exemple, un jugement d'ouverture de procédure de défaillance (dépôt de bilan d'une entreprise inscrite dans le cadre d'une procédure judiciaire) ne se solde pas forcément par une liquidation judiciaire. A contrario, toutes les cessations ne découlent pas d'une défaillance. Par exemple, un entrepreneur individuel peut cesser son activité pour cause de départ en retraite. (INSEE, 2019)

3.1.1.4 Caractérisation de la défaillance dans les thèses précédentes

En 2004, un état des lieux concernant la prévision de la faillite a été publié (Refait-Alexandre, 2004). Cet état des lieux commence par un rappel des cibles de chaque étude. L'auteur indique que la majorité des thèses caractérisent la défaillance par l'ouverture d'une procédure judiciaire. Cependant, certaines d'entre elles considèrent comme défaillante toute entreprise en état de cessation de paiement ou de défaut de paiement, cet état étant la préoccupation majeure des créanciers. La thèse donne la définition du risque de défaut de paiement, à savoir, le non-respect par le débiteur de ses obligations financières. Elle donne les exemples suivants : non-remboursement du capital ou non-versement des intérêts,

violation d'un covenant (clause imposant à l'emprunteur de respecter des ratios financiers visant à s'assurer du remboursement de la dette).

Néanmoins, elle ajoute que la notion de défaillance peut revêtir d'autres formes comme une « dégradation de la qualité de signature de l'entreprise : restructuration de la dette, diminution des dividendes versés, voire avis défavorable d'un audit ou encore détérioration du rating⁵ du débiteur. » (Refait-Alexandre, 2004)

Il est à noter que, d'une part, cette caractérisation de la défaillance par l'ouverture d'une procédure judiciaire peut faire l'objet d'adaptation au contexte et évoluer. D'autre part, cette caractérisation est antérieure à l'essor des procédures préventives et amiables depuis 2005, date à laquelle le législateur a ajouté la procédure de conciliation (Koehl, 2019).

Plus récemment, en Belgique, une étude dans le cadre d'un master a redéfini la caractérisation de la défaillance comme étant le reflet de l'activité. Trois états sont pris en considération dans cette étude des spin-offs universitaires⁶ belges : actif, racheté et en cessation d'activité (faillite) (Ferire, 2017).

La caractérisation de la défaillance des entreprises par l'activité n'a, jusqu'alors, pas été appliquée dans une étude de prédiction des défaillances des entreprises françaises.

3.1.2 Eventuels facteurs explicatifs de la défaillance

En 2017, un article de l'Université Paris Nanterre, initialement publié dans un ouvrage collectif en 2011, a décrit l'impact du poids de la routine et de l'absence des données qualitatives dans la sélection des facteurs potentiellement explicatifs lors des travaux de prédiction des défaillances (Lecointre, 2011) (Levratto, Le processus de défaillance des entreprises, 2017).

L'article catégorise en deux ensembles les facteurs contributifs de la défaillance. Le premier ensemble porte sur l'entreprise elle-même. Cet ensemble inclut aussi bien des variables financières issues des bilans ou comptes de résultats, que des variables non financières. Ces dernières sont relatives à sa structure, son organisation, son management, sa stratégie... Le deuxième ensemble de facteurs porte sur l'environnement de l'entreprise. Il peut s'agir de données générales tels que l'inflation, le taux d'intérêt, le taux de croissance (etc...), ou bien de données décrivant un secteur ou un marché.

⁵ Le rating est une appréciation donnée par une agence de notation financière. Elle reflète la perception du risque de non-remboursement d'une dette.

⁶ Une spin-off est une entreprise créée à partir des connaissances et technologies issues de la recherche. De cette manière, l'Université apporte directement ses activités scientifiques de haut niveau au sein de la société via l'exploitation industrielle et économique des connaissances et technologies qui y sont développées. (Définition de l'Université libre de Bruxelles)

Mais l'article établit le constat que l'étendue des facteurs contributifs cités précédemment est peu exploitée dans les études. En effet, les prédictions réalisées se basent sur une partie seulement de ces facteurs. La sélection opérée s'explique, bien sûr, par l'objectif de l'étude mais aussi par les contraintes rencontrées. D'abord, la manière de procéder n'évolue guère, ce qui se traduit par la répétition des méthodes déjà vérifiées. Une des causes de cette répétition est la contrainte de temps. Une autre contrainte concerne la disponibilité des données nécessaires. En effet, les données comptables présentaient l'avantage d'être obligatoirement publiées et donc d'obtention simple voire quasiment gratuite (INPI, 2017), deux avantages essentiels. A contrario, les données non-financières sont rares et payantes. De plus, la nouvelle option de confidentialité des comptes rend les données comptables de moins en moins disponibles. Tout ceci a pour conséquence de diminuer l'efficacité des modèles de prédiction des défaillances.

La revue de littérature qui suit confirme cette quasi-répétition des méthodes employées.

L'état des lieux publié en 2004 décrit les facteurs explicatifs à la défaillance exploités au fil des recherches dans différentes études jusqu'alors (Refait-Alexandre, 2004).

Il ressort de cet état des lieux que les données comptables sont toujours utilisées, parfois accompagnées de données boursières. En effet, les facteurs comptables et financiers prédominent depuis les études statistiques de Beaver en 1966, Altman en 1968 jusqu'aux études basées sur des techniques d'intelligence artificielle de Bardos et Zhu en 1997, Varetto en 1998.

Néanmoins, certaines études mentionnées, tentent de prendre en compte aussi le secteur économique en constituant un échantillon spécifique au secteur choisi (Trieschmann et Pinches en 1973 ; Altman en 1977, Altman et Loris en 1976 ; Calia et Ganuci en 1997).

D'autres études décrites exploitent des informations liées à l'organisation de l'entreprise ou à la nature de son financement, comme le nombre de banques auprès desquelles les entreprises sont endettées (Foglia, Marullo, & Marullo Reedtz, 1998).

Dans cette droite lignée, la Banque de France a construit, en 1995, une fonction score linéaire à partir des documents comptables. L'analyse financière de la défaillance pratiquée alors, identifie les ratios financiers discriminants : le risque patrimonial et le risque d'illiquidité, les risques liés à l'activité productive, à la rentabilité, au financement et à la gestion, ainsi qu'à l'évolution et aux structures de financement (Banque de France, 1995).

Plus récemment, une étude de prévision des défaillances, publiée en 2013, fait intervenir uniquement des ratios financiers déjà éprouvés (Ben Jabeur & Fahmi, 2013).

Les modèles anciens sont encore à l'honneur. C'est le cas, en 2017, dans le mémoire de validation de modèles prédictifs des défaillances au travers d'une analyse des spin-offs belges en faillite (Ferire, 2017). Son objectif est de vérifier que d'anciens modèles financiers de faillite sont toujours d'actualité, comme le ratio de Beaver, la méthode d'Altman, la méthode de Collongues et le scoring de Conan & Holder.

C'est toujours le cas des ratios financiers précédemment définis en 2017 (Barboza, Kimura, & Altman, 2017) et repris dans un article de blog en 2019 (Schwab, 2019).

Cependant, parmi les 11 ratios cités, se trouve un ratio non financier, à savoir, la tendance des effectifs de l'entreprise.

Enfin, en 2020, un point de vue est publié sur le sujet des défaillances des entreprises. Il donne un éclairage nouveau sur les facteurs contributifs suivants : l'âge, les effectifs, les dettes bancaires, les dettes non bancaires (fournisseurs, Urssaf, impôts) (France Stratégie Evaluer, Anticiper, Débattre, Proposer, 2020). La mesure de l'amélioration des modèles de prédiction est illustrée ci-dessous.

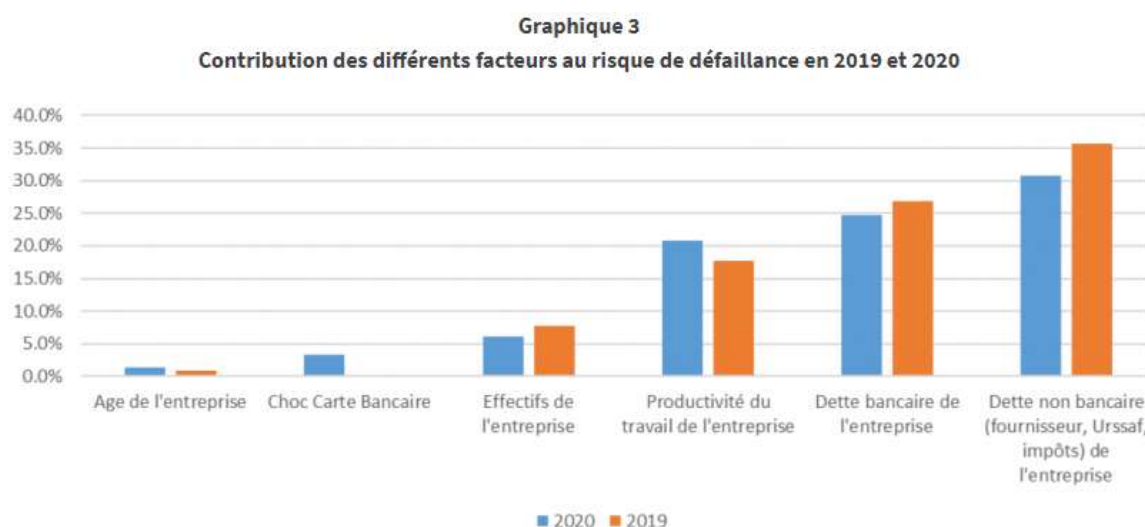


Figure 12 Mesure de la contribution des facteurs au risque de défaillance

Ce point de vue incite à revoir, avec succès, le choix des variables explicatives traditionnelles de la prédiction des défaillances.

3.1.3 « Trous de connaissances » en prédiction de défaillance

Peu, voire pas, de recherches ont pris en compte les évolutions récentes des lois relatives à l'aide aux entreprises en difficulté et à la publication de leur compte sous clause de confidentialité.

De même, peu de facteurs explicatifs non financiers ont été intégrés dans les prévisions récentes de défaillances des entreprises.

La présente problématique de recherche propose d'apporter une contribution sur ces thèmes.

3.1.4 Aspects « Métier » de la problématique de recherche

Une nouvelle problématique de recherche s'impose. Elle intègre une nouvelle caractérisation de la défaillance et la revue des facteurs contributifs potentiels au profit de l'ajout des variables non-financières.

L'approche empirique de ce mémoire propose de vérifier les hypothèses suivantes :

- La première hypothèse concerne la caractérisation de la défaillance avec la notion de cessation d'activité plus globale. La cessation d'activité peut-elle être un indicateur de la défaillance des entreprises ?
- La seconde hypothèse concerne la sélection a priori des facteurs contributifs. A côté des variables comptables et financières, des variables non financières peuvent-elles être ajoutées ? A minima, l'âge de l'entreprise, les effectifs, le secteur d'activité et les informations bancaires sont à prendre en compte.

Mais cette nouvelle problématique de recherche requiert une certaine disponibilité des données et de nouvelles techniques de prédiction. Ces dernières font l'objet du chapitre suivant.

3.2 Cadre théorique de la prédiction en classification

L'introduction a présenté, précédemment, l'historique de la conception de l'algorithme XGBoost et son utilisation lors de la détection des fraudes bancaires en 2019 (Metzler, 2019) et 2021 (Alfocea, 2021). La revue de littérature donne ici l'état de l'art de cet algorithme et de la classification.

3.2.1 Etat de l'art en classification

Pour rappel, XGBoost est un algorithme de type Gradient Boosting. Ce type d'algorithme est lui-même issu des algorithmes de Boosting.

Les contributions du Boosting au Machine Learning ont été décrites dans une thèse de 2006 (Suchier, 2006).

L'approche standard du Boosting consiste à apprendre successivement des hypothèses dites faibles et à les combiner en une hypothèse dite forte. Chaque itération corrige la précédente par un jeu de pondération des exemples mal classés. Les exemples d'apprentissage sont pondérés de la façon suivante : les exemples mal classés sont augmentés de façon exponentielle, alors que les exemples bien classés sont diminués. La repondération est fonction d'un coefficient issu de l'erreur calculée en apprentissage. Ce processus est répété plusieurs fois. L'hypothèse finale est la combinaison de toutes les hypothèses faibles ainsi construites, pondérées par les coefficients décrits précédemment. Les trois principales études des performances de cet algorithme (Freund & Schapire, 1997 ; Schapire et al., 1997 ; Schapire & Singer, 1999) sont mentionnées dans cette thèse. L'algorithme converge bien en continuant de faire diminuer l'erreur en phase de généralisation avec un nombre important d'itérations, ce qui n'est pas habituel. En effet, les arbres de décision ont la réputation de « surapprendre » les exemples en apprentissage ce qui les conduit, en phase de généralisation, à stopper la baisse de l'erreur de prédiction assez rapidement et à l'augmenter au fur et à mesure que le nombre d'itérations se poursuit.

De plus, l'extension du Boosting comme méthode de descente du gradient y est décrite. En 2001, un article rappelle la notion de descente de gradient comme étant la recherche de descente de la pente la plus raide pour déterminer le minimum d'une fonction (Friedman, 2001). La descente de la pente se fait le long de la courbe de la fonction suivant la pente la

plus forte avec un pas de progression de longueur donnée. La valeur courante x_t , qui indique les variations de F autour de x_{t-1} , est calculée à chaque pas de la façon suivante :

$$x_t = x_{t-1} + \epsilon \nabla F(x_{t-1})$$

où ϵ caractérise la longueur du pas, et $\nabla F(x_{t-1})$ est le gradient de F en x_{t-1} .

Il décrit ensuite l'utilisation de la descente de gradient dans le cadre des algorithmes de type Gradient Boosting. Ces algorithmes consistent en un assemblage d'arbres de décision successifs appliquant la technique de la descente de gradient pour corriger l'erreur de prédiction de l'arbre précédent à chaque itération, mesurée par la fonction de coût. L'erreur, appelée aussi résidu, se trouve rapidement minimisée ainsi. Les expressions mathématiques figurent ci-après. A chaque itération, un modèle faible F_k est appris en utilisant l'erreur obtenue par la combinaison linéaire du modèle précédent. La combinaison linéaire F_k à l'instant k est définie comme suit :

$$F_k = F_{k-1} + \alpha_k f_k$$

Où F_{k-1} est la combinaison linéaire des $k-1$ premiers modèles et α_k le poids donné au $k^{\text{ème}}$ modèle faible. Les modèles faibles sont entraînés avec la mesure de l'erreur résiduelle suivante du modèle courant :

$$r_i = y_i - F_{t-1}(x_i)$$

Ce résidu est obtenu avec le gradient négatif, $-g_t$, de la fonction de perte correspondant à la prédiction courante $F_{t-1}(x_i)$:

$$r_i = g_k(x_i) = - \left[\frac{\partial L(y_i, F_{k-1}(x_i))}{\partial F_{k-1}(x_i)} \right]$$

Une fois les résidus r_i calculés, le problème d'optimisation est résolu de la façon suivante :

$$(f_k, \alpha_k) = \underset{\alpha, f}{\operatorname{argmin}} \sum_{i=1}^m (r_i - \alpha f(x_i))^2$$

A l'itération suivante, la combinaison F_k , énoncée précédemment, permet le calcul du modèle suivant, et ainsi de suite.

XGBoost est une implémentation particulière d'algorithme de Gradient Boosting conçue pour améliorer ses performances et sa vitesse. Il a été présenté en 2016 à la communauté internationale et a remporté un vif succès (Chen & Guestrin, 2016). Il apporte un ensemble d'améliorations comme : la prévention du surapprentissage, la robustesse aux valeurs manquantes, la parallélisation des arbres, l'amélioration des performances grâce à l'élagage des arbres, etc...

En particulier, la gestion des valeurs manquantes apporte une solution au problème répandu de la rareté de la donnée. Elle peut être due à l'absence de donnée, à la fréquente valorisation à zéro en statistiques ou à un effet indésirable de l'encodage des valeurs en pré-traitement avec les fonctions de type one hot encoding. Cette gestion consiste à ajouter une

direction par défaut à chaque nœud de l'arbre. Quand une valeur manquante est rencontrée, l'exemple est classé dans la direction par défaut. En réalité, il y a deux choix de direction par défaut à chaque branche. Les deux choix sont évalués en parallèle : le choix de la direction résultant dans un gain maximum est retenu, ainsi que la séparation des sous-échantillons correspondante.

De plus, l'hyperparamètre « lambda » diminue le surapprentissage en pénalisant les modèles complexes qui se sur-adaptent trop à l'échantillon d'entraînement.

Enfin, la technique de l'élagage des arbres (Tree Truning) améliore les performances grâce à son hyperparamètre « max_depth ». Les algorithmes de type Gradient Boosting sont par nature gourmands du fait de leur critère d'arrêt de division des exemples. En effet, le critère consiste à atteindre une valeur de fonction de coût négative : l'erreur résiduelle a cessé de diminuer et elle recommence à augmenter. L'hyperparamètre « max_depth » limite cet effet en élaguant les arbres en profondeur et améliore les performances de calcul.

Rapidement, XGBoost a fait preuve de sa supériorité. Une publication de 2019 le démontre dans le domaine de la classification d'images, où XGBoost obtient de meilleurs résultats qu'un réseau de neurones (Memon, Patel, & Patel, 2019).

Une utilisation inédite de l'algorithme XGBoost mérite d'être signalée (Deng, Li, Deng, & Wang, 2021). Dans le domaine de la prédiction des cancers, un modèle à deux étages a été conçu pour d'abord sélectionner les variables importantes, puis effectuer la prédiction en elle-même. En effet, XGBoost permet d'obtenir automatiquement l'importance des variables pour filtrer celles-ci. Les plus significatives sont ensuite utilisées dans le second algorithme.

Enfin, un mémoire présente les avantages de l'utilisation de l'algorithme XGBoost dans la prédiction des sinistres en Garantie contre les accidents de la vie, en comparaison des méthodes de provisionnement traditionnelles (Ottou, 2017). Les arguments donnés en la faveur de XGBoost sont les suivants : la restitution des variables les plus discriminantes du modèle construit, la prise en compte des variables liées, l'efficacité de traitement des problèmes non linéaires, i.e. lorsque la variable à prédire ne dépend pas linéairement des variables explicatives.

Enfin, il s'avère que l'application de la bibliothèque open source XGBoost à la prédiction de défaillance des entreprises n'a pas encore fait l'objet de publication.

3.2.2 Classification dans un contexte de déséquilibre de classe

3.2.2.1 Affichage des résultats de la classification avec la matrice de confusion

La matrice de confusion est un tableau faisant apparaître les dénombrements importants de la prédiction réalisées. Ces dénombrements sont catégorisés suivant leur vrai caractère et leur caractère prédit.

Par convention, un exemple de la classe minoritaire (positif), prédit positif par le modèle, est appelé True Positive (TP). Si l'exemple est mal classé par le modèle, il est appelé False Negative (FN). Suivant la même logique pour la classe majoritaire (négatif), un exemple bien classé est appelé True Negative (TN) et mal classé False Positive (FP).

Prédiction Positif / Négatif	Vrais Négatifs	Faux Positifs
	Faux Négatifs	Vrais Positifs
	Négatifs / Positifs	Réalité

3.2.2.2 Mesure du déséquilibre avec le Ratio de déséquilibre

Dans le monde réel, la classe à prédire apparaît parfois de façon très faible dans l'ensemble de données. Celui-ci présente un « déséquilibre de classe » lorsque le déséquilibre dépasse 10% / 90%. Le déséquilibre est mesuré par le Ratio de déséquilibre (ou Imbalance ratio en anglais) calculé de la façon suivante :

$$\text{Ratio de déséquilibre} = \text{Imbalance ratio} = \frac{\text{Nombre d'observations majoritaires}}{\text{Nombre d'observations minoritaires}}$$

Le ratio vaut 1 lorsque les données sont parfaitement équilibrées.

Par convention, les observations positives, i.e. comportant le signal recherché, appartiennent à la classe minoritaire et les observations négatives à la classe majoritaire. Dans le cas d'une détection de fraude : les opérations valides appartiennent à la classe majoritaire et les opérations frauduleuses appartiennent à la classe minoritaire.

En conséquence, le Ratio de déséquilibre est donc supérieur ou égal à 1.

3.2.2.3 Problèmes posés pour le modèle par arbre

Dans le modèle par arbre, le déséquilibre de classe génère une augmentation de l'erreur de classement et une variance élevée des résultats (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). En effet, les arbres ont tendance à surapprendre : ils localisent les quelques observations positives dans les larges espaces d'observations négatives et ainsi réduisent leur biais en localisant strictement les observations positives sur l'ensemble de données d'entraînement.

3.2.2.3.1 Surapprentissage

Mais, la localisation des observations positives peut varier légèrement sur l'ensemble de données de test. Le modèle se trompe alors : les observations positives ne sont plus localisées et le modèle génère des faux négatifs. Par ailleurs, des faux positifs sont générés dans les régions surappprises. En conséquence, les performances, mesurant les résultats, chutent : une variance élevée caractéristique du surapprentissage (overfitting) est constatée.

3.2.2.3.2 Apprentissage régularisé mais sous-optimum

Le problème pourrait trouver une solution dans la régularisation des arbres en ajustant certains paramètres. Dans le cas de XGBoost, les paramètres liés au renforcement des arbres et les hyperparamètres du Gradient Boosting sont concernés. Il s'agit alors, pour les premiers paramètres de réduire la profondeur maximale de l'arbre, d'augmenter le nombre minimum d'échantillons requis pour la division d'un nœud interne, l'augmentation minimum de la pureté pour la division du nœud, le maximum de feuilles par nœud, etc... Pour les hyperparamètres, il s'agit d'ajuster le ratio de sous-échantillonnage d'entraînement, le pas d'apprentissage...

Mais, dans le cas d'un signal complexe, difficile à apprendre, cela conduit à diminuer les performances du modèle : la variance est diminuée au prix d'une simplification trop grande, ce qui est sous-optimum.

3.2.2.4 Mesure de la performance adaptée au déséquilibre

3.2.2.4.1 Mesure classique de la performance

La mesure classique de la justesse (accuracy) ne convient pas dans un contexte de déséquilibre de classe. Savoir qu'une erreur de classification a été commise ne suffit pas. Il faut aussi savoir dans quelle classe l'erreur a été commise.

Les formules des mesures de performance se basent sur les dénombrements : TP, FN, TN et FP, qui apparaissent sur la matrice de confusion. La plus classique est la justesse :

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+TN+FP}$$

Son corollaire est le taux d'erreur.

$$\text{Error rate} = \frac{FP+F}{TP+FN+TN+F}$$

Néanmoins, si l'ensemble comporte 1% de données positives, un modèle qui prédit toutes les données négatives, a une mesure d'accuracy élevée de 99%. Mais, toutes les données intéressantes, à savoir appartenant à la classe minoritaire, ont été mal prédites.

D'autres formules de performance sont plus pertinentes, comme le Rappel (Recall ou Sensitivity).

$$\text{Recall} = \frac{TP}{TP+FN}$$

Le rappel mesure le pourcentage d'exemples de la classe d'intérêt minoritaire que le modèle est capable de prédire.

Une autre mesure est la Précision.

$$\text{Precision} = \frac{TP}{TP+F}$$

La précision mesure comment le modèle réussit à prédire un exemple positif.

3.2.2.4.2 Mesures statistiques

D'autres mesures sont utilisées en statistiques, notamment le taux de sous-estimation (False Negative Rate).

3.2.2.4.2.1 True Positif Rate ou sensibilité ou encore rappel en français

Le True Positive Rate est le taux de vrais positifs parmi tous les positifs. Il est calculé de la façon suivante :

$$TPR = \frac{TP}{P}$$

Avec $P = TP + FN$ (P étant la somme des positifs)

Le TPR est aussi appelé sensibilité.

3.2.2.4.2.2 Sous-estimation (ou False Negative Rate en anglais)

Le taux de sous-estimation est le taux de faux négatifs parmi tous les positifs. Il est calculé de la façon suivante :

$$AS = FNR = \frac{FN}{P} = \frac{FN}{TP + FN} = \frac{TP + FN - TP}{TP + FN} = 1 - \frac{TP}{TP + FN} = 1 - TPR$$

Toujours avec $P = TP + FN$ (P étant la somme des positifs)

Le taux de sous-estimation porte aussi le nom de miss rate en anglais (ou de FNR).

Le taux de sous-estimation mesure l'erreur de type II en statistiques.

Une erreur de type II survient dans un test d'hypothèse statistique lorsque l'hypothèse nulle est acceptée par erreur. Elle représente l'échec de détection d'un effet positif alors qu'il existe.

NB : la spécificité représente la capacité d'un test à donner un résultat négatif lorsque l'hypothèse n'est pas vérifiée. Cette notion est d'une importance majeure en épidémiologie et en théorie de la détection du signal (Wikipédia, s.d.)

3.2.2.4.2.3 False Positif Rate

Le False Positive Rate est le taux de faux positifs parmi tous les négatifs. Il est calculé de la façon suivante :

$$FPR = \frac{FP}{N}$$

Avec $N = TN + FP$ (N étant la somme des négatifs)

Il est utilisé dans la courbe ROC qui est décrite plus loin.

3.2.2.4.3 F-Mesure

Dans le contexte de la prédiction de fraude, les mesures du rappel et de la précision sont toutes les deux importantes. C'est pourquoi la F-Mesure, qui les prend en compte ensemble a été retenue. Elle est définie comme une moyenne harmonique du rappel et de la précision dépendant d'un paramètre β .

$$F_{\beta} = \frac{(1+\beta^2)Precision*Recall}{\beta^2*Precision+Recall}.$$

En modifiant la valeur de β , l'utilisateur peut modifier l'importance du Rappel ou de la Précision dans la F-Mesure. La F-Mesure est aussi appelée F-Score.

Une thèse de 2019 présente une application de la F-Mesure (Metzler, 2019). Elle brosse un panorama des solutions possibles pour gérer les classes déséquilibrées dans le cadre de la détection de fraude bancaire. Dans un chapitre dédié aux méthodes sensibles aux coûts, elle détaille notamment une solution d'optimisation de la F-Mesure : l'étude propose d'assigner à chaque classe un paramètre de meilleur coût pour maximiser la métrique de performance F-Mesure calculée pour les besoins.

3.2.2.4.4 Courbe ROC et ROC AUC score

Le calcul unique d'une mesure est néanmoins insuffisant.

La courbe ROC (Receiver Operating Characteristic) fournit une vue synthétique des performances d'un modèle tout en faisant varier le seuil de classification.

En effet, un seuillage est appliqué au seuil de probabilité entre les classes positives et négatives, qui par défaut pour tout classificateur est fixé à 0,5, à mi-chemin entre chaque résultat (0 et 1).

La modification du seuil de classification modifiera l'équilibre des prédictions vers l'amélioration du True Positive Rate au détriment du False Positive Rate, ou le cas inverse. Ce sujet est détaillé plus loin au chapitre Ajustement du seuil.

Un compromis est à déterminer entre ces deux taux.

« En évaluant les vrais positifs et les faux positifs pour différentes valeurs de seuil, une courbe peut être tracée qui s'étend du bas à gauche vers le haut à droite... Cette courbe est appelée courbe ROC. L'avantage de cette courbe est de ne pas présenter de biais en faveur des modèles qui fonctionnent bien sur la classe minoritaire au détriment de la classe majoritaire » (He & Ma, 2013).

Le score ROC AUC (Area Under Curve) est la mesure de l'aire sous la courbe ROC. Il donne un score unique pour un modèle de classificateur portant sur toutes ses valeurs de seuil.

« Ce score a une valeur comprise entre 0 et 1. Il peut être interprété comme la probabilité que les scores donnés par un classificateur classent une instance positive choisie au hasard plus haut qu'une instance négative choisie au hasard. » (Fernandez, et al., 2018)

Néanmoins, le score ROC AUC peut être trompeur. En effet, dans le cas d'une classification déséquilibrée avec un biais sévère et quelques exemples de la classe minoritaire, un petit nombre de prédictions correctes ou incorrectes seulement peut entraîner une modification importante de la courbe ROC ou du score ROC AUC (Fernandez, et al., 2018).

L'alternative est la courbe Précision Rappel.

3.2.2.4.5 Courbe Précision Rappel et score AUC PR

Pour mémoire, la précision est une métrique qui quantifie le nombre de prédictions positives correctes faites et le rappel quantifie le nombre de prédictions positives correctes faites à partir de toutes les prédictions positives qui auraient pu être faites. Ces deux métriques sont donc centrées sur la classe positive (minoritaire) et ne sont pas concernées par la classe négative (majoritaire).

La courbe Précision Rappel est le tracé de la précision par rapport au rappel pour différents seuils de probabilités. Ce tracé est souvent accompagné du tracé d'une courbe correspondant à un classifieur « No Skill » classifiant les observations positives au hasard. Cette courbe « No Skill » est alors une ligne égale à la proportion de la classe positive dans le dataset.

« Les courbes de rappel de précision (courbes PR) sont recommandées pour les domaines fortement asymétriques où les courbes ROC peuvent fournir une vue excessivement optimiste de la performance » (Branco, Torgo, & Ribeiro, 2015)

Le score AUC PR est la mesure de l'aire sous la courbe Precision Recall. De même que le score ROC AUC, le score AUC PR donne une synthèse de la courbe pour les différentes valeurs de seuils des scores. Il vaut 1 dans le cas d'un classifieur parfait. Dans le cas du classifieur « No Skill », l'AUC PR est égale à la proportion de la classe positive dans le dataset.

Une façon de calculer l'AUC PR est de chercher la valeur de l'AP, la précision moyenne, avec la fonction `sklearn.metrics.average_precision_score`. L'AP est une sorte de moyenne pondérée de la précision à travers tous les seuils (Steen, 2020).

En conclusion, le score AUC PR permet de mieux comparer les performances des modèles dans un contexte de déséquilibre sévère aggravé par le manque d'exemples de la classe minoritaire (Brownlee, 2020).

3.2.2.5 Correction ou non du déséquilibre de classe

Le déséquilibre de classe n'est pas obligatoirement à corriger à l'équilibre 50 % - 50 %. En effet, l'optimum dépend du classifieur et de l'ensemble des données. Un ordre de grandeur admis est de 5 – 10 % de classe minoritaire pour les modèles à arbres. La détermination du ratio de déséquilibre tolérable est toutefois empirique.

3.2.2.6 Trois niveaux de solutions de correction du déséquilibre

3.2.2.6.1 Les solutions de niveau Data

Ces solutions consistent à modifier l'échantillonnage :

- Le sous-échantillonnage aléatoire (random undersampling) :
Il porte sur les observations majoritaires. Des observations majoritaires sont retirées soit de manière globale, soit de manière spécifique.
- Le sur-échantillonnage aléatoire (random oversampling)
Il porte sur les observations minoritaires. Des individus minoritaires sont rajoutés (« clonés »).
- Le sur-échantillonnage synthétique (SMOTE pour Synthetic Minority Oversampling)
Il porte aussi sur les observations minoritaires. Des observations minoritaires ressemblantes mais distinctes sont ajoutées.

La technique SMOTE a été récemment appliquée dans une étude de la prédiction des défaillances bancaires comparant l'approche statistique et Machine Learning (Durand & Le Quang, 2021).

3.2.2.6.2 Les solutions de niveau algorithmique

- Le ré-échantillonnage interne des arbres dans les méthodes de Boosting ou de Bagging
Cette solution est codée dans les `BalancedBaggingClassifier` ou `RUSBoostClassifier`.
Un exemple de `BalancedBaggingClassifier`, à savoir le `BalancedBaggingXGBoost`, a été comparé à différents modèles de prédiction de la fraude bancaire (Alfocea, 2021).
- La surpondération globale des observations minoritaires dans l'apprentissage
 - Via le gradient (Boosting)
C'est le rôle de l'hyperparamètre `scale_pos_weight` dans XGBoost, qui permet de donner davantage de poids aux observations minoritaires. Le choix de cet hyperparamètre est abordé dans l'approche empirique suivante.
 - Via le Gini (Bagging)

3.2.2.6.3 Les solutions sensibles aux coûts

Le cost-sensitive learning (apprentissage sensible aux coûts) consiste à surpondérer les faux négatifs (les positifs qui sont prédits négatifs) par rapport aux faux positifs durant la phase d'apprentissage. Cette méthode est à bien distinguer des méthodes précédentes de surpondération globale. En effet, le cost-sensitive learning s'attache à surpondérer uniquement les positifs mal-classés. Pour comparaison, les méthodes globales surpondèrent tous les positifs.

Plusieurs exemples de cost-sensitive learning, à savoir le cost-sensitive SVC et le cost-sensitive XGBoost, ont, eux aussi, été comparés dans le cadre de la prédiction de fraude bancaire (Alfocea, 2021).

3.2.3 Ajustement du seuil de score de classification

Pour mémoire, le seuil de score de classification permet de trier les observations positives des observation négatives. Mais le seuil défini par défaut peut entraîner de mauvaises performances dans le cas (grave) de déséquilibre de classe. Une approche simple et directe consiste à régler le seuil utilisé pour diviser les probabilités d'appartenance aux classes positives et négatives.

La littérature comprend des publications qui étudient l'ajustement d'un seuil de score de classification à des fins d'amélioration des performances de la prédiction.

En particulier, les méthodes d'ajustement et de réglage du seuil dans le cadre d'un déséquilibre de classe ont été détaillées en 2020 (Brownlee, 2020).

Le chapitre suivant rappelle ce qu'est le score de classification et la mesure de son écart à la réalité avec le Score Log Loss.

3.2.3.1 *Méthode de score des probabilités*

3.2.3.1.1 La probabilité d'appartenance à une classe

La modélisation prédictive de classification implique la prédiction d'une étiquette de classe pour des exemples. Mais certains problèmes nécessitent la prédiction d'une probabilité d'appartenance à une classe. Pour ces problèmes, les étiquettes de classe nettes ne sont pas souhaitables ; la probabilité que chaque exemple appartienne à chaque classe est plus adaptée. En effet, La probabilité synthétise une vraisemblance, ou une incertitude, qu'une observation appartienne à une classe. La probabilité est donc plus nuancée : elle offre une information de granularité supplémentaire.

En pratique, un jeu de données n'aura pas de probabilités cibles. Au lieu de cela, il aura des étiquettes de classe. Concrètement, on passe de l'étiquette de classe à la probabilité de la façon suivante : lorsqu'un exemple possède l'étiquette de classe 1, alors la probabilité des étiquettes de classe 0 et 1 sera, respectivement, 0 et 1.

C'est la régression logistique.

3.2.3.1.2 Objectifs des métriques de probabilités

Les métriques de probabilité sont spécifiquement conçues pour mesurer les performances d'un modèle de classificateur en utilisant les probabilités prédites au lieu d'étiquettes de classe précises. Ce sont généralement des scores qui fournissent une valeur unique qui peut être utilisée pour comparer différents modèles. Ces métriques résumeront dans quelle mesure la distribution prédite de l'appartenance à une classe correspond à la distribution de probabilité de la classe connue.

Ce centrage sur les probabilités prédites implique que les étiquettes de classe nettes prédites par un modèle sont ignorées dans la mesure de performance. En effet, un modèle qui prédit des probabilités peut sembler avoir des performances médiocres lorsqu'il est évalué en fonction de ses étiquettes de classe nettes, telles que l'utilisation de la précision ou d'un score similaire. C'est pourquoi, bien que les probabilités prédites puissent montrer

de bonnes performances, elles doivent être interprétées avec un seuil approprié avant d'être converties en étiquettes de classe nettes.

Une métrique de probabilités est généralement calculée pour chaque exemple, puis moyennée sur tous les exemples de l'ensemble de données d'apprentissage.

Il existe deux mesures courantes pour évaluer les probabilités prédites : la Log Loss et le Score Brier. La Log Loss est présentée ci-après.

3.2.3.1.3 Score Log Loss

La perte logarithmique, ou Log Loss en anglais, est une fonction de perte connue pour entraîner l'algorithme de classification par régression logistique.

La fonction Log Loss se base sur le calcul d'une vraisemblance des prédictions de probabilité faites par le modèle de classification binaire. Plus particulièrement, il s'agit de la classification par régression logistique, mais cette fonction peut être utilisée par d'autres modèles, tels que les réseaux de neurones, et est connue sous d'autres noms, tels que l'entropie croisée (Binary Cross-Entropy).

La perte peut être calculée en utilisant les probabilités attendues pour chaque classe et le logarithme népérien des probabilités prédites pour chaque classe. Par exemple, en classification binaire :

$$Log Loss = -\frac{1}{N} \sum_{i=1}^N y_i * \log(p(y_i)) + (1 - y_i) * \log(1 - p(y_i))$$

Où y_i est l'étiquette de classe (0 pour une observation négative ou 1 pour une observation positive), $p(y_i)$ la probabilité prédite de l'observation et N le nombre d'observations.

La meilleure valeur de la Log Loss est une perte de 0. La Log Loss prend ses valeurs de 0 à $-\infty$ pour ses pires valeurs.

C'est la mesure utilisée par défaut par l'algorithme XGBoost dans le cas d'une classification par régression logistique.

D'autres fonctions de perte existent comme le Brier Score qui est basé sur le calcul de la MSE (Mean Square Error) des probabilités prédites.

3.2.3.2 Ajustement du seuil de probabilité

La modification du seuil de classification modifie l'équilibre des prédictions au profit de l'amélioration du True Positive Rate et au détriment du False Positive Rate, ou le sens inverse.

Le seuil optimum peut être localisé et calculé automatiquement lorsqu'il est basé sur les courbes ROC ou Précision Rappel. Dans d'autres cas, une recherche en grille (GridSearchCV) permet de régler le seuil et de définir sa valeur.

3.2.3.2.1 Conversion des probabilités à l'étiquette de classe

La classification binaire nécessite une tâche de prédiction des étiquettes de classe nettes.

Ceci implique qu'une prédiction d'appartenance à une classe doit être convertie en étiquette de classe nette. La décision de conversion dans une classe ou un autre est régie par un « paramètre » appelé seuil (threshold).

La valeur par défaut du seuil est de 0,5 pour les probabilités prédites normalisées ou les scores compris entre 0 et 1. Les valeurs de score de probabilité inférieures au seuil sont assignées à la classe négative et les valeurs supérieures ou égale au seuil sont assignées à la classe positive.

Mais, le seuil par défaut peut ne pas être adapté pour différentes raisons :

- Les probabilités prédites ne sont pas calibrées,
- La métrique utilisée pour entraîner le modèle n'est pas la métrique utilisée pour évaluer le modèle final,
- Les classes sont sévèrement déséquilibrées,
- Le coût d'un type d'erreur de classification est plus important qu'un autre. Par exemple, le coût d'un False Negative est plus important que le coût d'un False Positive.

3.2.3.2.2 Ajustement du seuil dans le cas d'un déséquilibre de classe

Il existe différentes solutions de correction d'un déséquilibre de classe sévère. La plus simple, et néanmoins efficace, peut être d'ajuster le seuil de classification.

Pour cela, différentes méthodes d'ajustement et de réglage sont possibles :

- La courbe ROC peut être analysée pour déterminer le seuil conduisant au meilleur équilibre entre True Positive Rate et False Positive Rate,
- La courbe Précision-Rappel peut être analysée pour déterminer le seuil conduisant au meilleur équilibre entre mesures de Précision et de Rappel,
- La métrique basée sur les probabilités utilisée pour entraîner, évaluer ou comparer le modèle, telle que la Log Loss, peut globalement servir à déterminer le seuil optimum
- La matrice de coûts, associant des coûts différents aux différents types d'erreurs de classification, peut servir à déterminer le meilleur compromis pour les prédictions.

Le processus d'ajustement du seuil est commun à toutes ces méthodes :

- Ajuster le modèle sur un ensemble de données d'apprentissage et à faire des prédictions sur un ensemble de données de test.
- Eventuellement, transformer en probabilités normalisées les prédictions qui se présentent sous la forme de probabilités de scores non normalisées.
- Essayer différentes valeurs de seuil et évaluer les étiquettes nettes résultantes à l'aide d'une métrique d'évaluation choisie.
- Adopter le seuil qui atteint la meilleure métrique d'évaluation, lors de la réalisation de prédictions sur de nouvelles données à l'avenir.

Deux méthodes d'ajustement et de réglage du seuil sont décrites ci-après.

3.2.3.2.3 Seuil optimum pour la courbe ROC ou la courbe Précision Rappel

Les courbes ROC et Précision Rappel permettent d'évaluer l'ensemble des prédictions de probabilité effectuées par un modèle sur l'ensemble de données de test.

La courbe ROC représente graphiquement le compromis entre le True Positive Rate et le False Positive Rate. Un modèle parfait se situe au point en haut à gauche, pour lequel le True Positive Rate est maximum et le False Positive Rate minimum. Le seuil optimum pour la courbe ROC est le seuil le plus proche de ce point. Il peut être défini grâce au calcul de la Moyenne Géométrique, basée sur le True Positive Rate et le False Positive Rate.

La courbe Précision Rappel est centrée, elle, uniquement sur les performances du classifieur sur la classe positive. Cette courbe représente graphiquement le compromis entre la mesure de la Précision, qui décrit comment le modèle prédit bien la classe positive et celle du Rappel qui décrit le pourcentage de positif que le modèle est capable de prédire correctement. Un modèle dit sans compétence (No-Skill) est représenté par une ligne avec une Précision égale au taux d'exemples positifs dans l'ensemble des données ($TP / (TP + TN)$). Le modèle parfait se situe au point en haut à droite, pour lequel la Précision et le Rappel sont maximum. Le seuil optimum pour la courbe Précision Rappel est le seuil le plus proche de ce point. Sa détermination revient à optimiser le F1-Score qui synthétise la moyenne harmonique des deux mesures : Précision et Rappel.

$$F1 - Score = \frac{2 * Précision * Rappel}{Précision + Rappel}$$

Dans les 2 cas, les seuils sont calculés à partir des mesures de performance issues du modèle.

3.2.3.2.4 Réglage du seuil optimum

Le réglage du seuil optimum peut s'effectuer aussi à partir des probabilités calculées par le modèle directement.

Dans ce cas, chaque valeur de seuil conduit à une évaluation de sa performance (avec le F1-Score par exemple). Ensuite, la meilleure valeur de la performance permet de déterminer le seuil optimum.

3.2.3.3 Application de l'ajustement du seuil dans la littérature

Cette application a été décrite dans une thèse de Machine Learning dans le domaine des assurances. (Ly, 2019). Elle compare les performances de différents modèles pour deux choix de seuil de score distincts : le seuil par défaut 0,5 et un seuil optimum calculé d'après la courbe ROC.

	0.5 cutoff		cutoff optimal	
	spécificité	sensitivité	spécificité	sensitivité
logit	0.9967	0.0057	0.7278	0.6351
bagging	0.9779	0.0661	0.6443	0.7069
random forest	0.9892	0.0316	0.6345	0.6954
boosting	0.9987	0.0000	0.6860	0.7385

Table 14.: Achat d'assurance: sensibilité au choix du seuil de classification.

Figure 13 exemple d'ajustement du seuil de classification en assurance

3.2.4 Différentes métriques de l'importance relative des variables

3.2.4.1 Interprétabilité

L'interprétabilité donne les réponses à la question "comment" un algorithme prend une décision (quels calculs, quelles données internes...). Une décision algorithmique est dite interprétable s'il est possible d'identifier les caractéristiques ou variables qui participent le plus à la décision, voire d'en quantifier l'importance. Cette interprétabilité est souhaitable voire nécessaire dans certains contextes, réglementaire notamment.

Elle favorise l'adoption des algorithmes et est une des raisons du succès des arbres de décision (Wikistat.fr, 2016). La mesure des critères d'importance est fournie dans certaines bibliothèques de Machine Learning comme celle de XGBoost.

Différentes métriques y sont fournies. Deux d'entre-elles sont décrites ci-dessous.

Tout d'abord, le « gain » peut être paramétré dans la fonction « get_score » de la library XGBoost pour les arbres de décision (par défaut « weight »). Il représente le « gain » moyen à travers toutes les divisions de la variable dans l'arbre.

Il s'agit de la moyenne de la réduction de la fonction coût (pour les données d'entraînement), quand une variable est utilisée pour une division au niveau d'un nœud (la Log Loss dans cette approche empirique).

Ensuite, la métrique, par défaut, « weight » de la fonction « plot_importance » représente le nombre de fois où la variable apparaît dans l'arbre.

L'importance est calculée pour un seul arbre de décision par le montant de la mesure de performance que chaque point de division de la variable améliore, pondéré par le nombre d'observations dont le nœud est responsable. La mesure de performance peut être la pureté (indice de Gini) utilisée pour sélectionner les points de partage ou une autre fonction d'erreur plus spécifique. Dans notre approche empirique, il s'agit de la fonction de coût Log Loss ou de la métrique « weight ».

Les importances des variables sont ensuite moyennées sur tous les arbres de décision du modèle.

Mais, les deux mesures d'importances donnent des résultats différents ! Toutefois, l'importance « gain » peut paraître plus pertinente (aquila data enabler, 2020).

Des préconisations émergent dans la littérature en faveur de l'utilisation de mesures alternatives, comme la permutation importance ou les valeurs de SHAP (SHapley Additive exPlanations).

Par ailleurs, de sérieuses mises en garde sont émises à l'encontre des corrélations des variables qui affectent la mesure de l'importance de celles-ci.

Cependant, l'importance des variables permet de mieux comprendre un modèle et de le valider auprès du « Métier ». Son étude est tout aussi nécessaire que l'analyse des données.

Il est à noter que l'interprétabilité des algorithmes de Machine Learning est une question ouverte et qu'elle fait toujours l'objet de recherches.

3.2.4.2 Explicabilité

L'explicabilité diffère de l'interprétabilité.

Une décision algorithmique est dite explicable s'il est possible d'en rendre compte explicitement à partir de données et caractéristiques connues de la situation.

L'explicabilité de la régression linéaire et de la régression logistique est comparée ci-dessous.

La régression linéaire donne une équation de la forme $Y = mX + C$, de degré 1.

Or, la régression logistique donne une équation moins triviale qui est de la forme exponentielle $Y = e^X * e^{-X}$

Les valeurs de SHAP peuvent aussi donner une explication de la prédiction du modèle sous la forme linéaire de M variables :

$$f(x) = y_{pred} = \varphi_0 + \sum_{i=1}^M \varphi_i * z'_i$$

Avec y_{pred} la valeur prédite du modèle pour cette exemple, φ_0 est la valeur de base du modèle, $z' \in \{0,1\}^M$ quand la variable i est présente $z'_i = 1$, $z'_i = 0$ sinon quand la variable est inconnue (George, Dehoux, & Liao, 2019).

Les valeurs de SHAP ont pour expression générale :

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

Avec M le nombre de variables, S un ensemble de variables, $f_x(S) = E[f(x)|x_S]$, i est la $i^{\text{ème}}$ variable.

3.2.5 « Trous de connaissances » en classification

La recherche renouvelle sans cesse les algorithmes de classification binaire, mais certains champs restent à explorer.

Le déséquilibre de classe a déjà fait l'objet de publication. Néanmoins, comment définir le ratio de déséquilibre tolérable ?

Certains ratios de performance ne sont pas exploités dans l'évaluation des modèles de Machine Learning. C'est le cas du taux de sous-estimation (ou FNR en anglais). Or, ce taux peut être très pénalisant pour l'adoption d'un modèle. Comment intégrer le taux de sous-estimation dans l'évaluation du modèle ?

Seules certaines méthodes d'ajustement du seuil de classification ont été publiées. La diminution du taux de sous-estimation n'a pas été leur objectif. Le taux de sous-estimation peut-il motiver l'ajustement du seuil de classification avec efficacité ?

La correction du déséquilibre de classe peut s'effectuer par le biais de nombreuses combinaisons de solutions. De nouvelles combinaisons de solutions de correction de déséquilibre de classe sont-elles possibles ?

La présente problématique de recherche propose d'apporter une réponse à ces différentes questions.

3.2.6 Aspects « Machine Learning » de la problématique de recherche

Les études publiées dans la littérature sont diverses mais ne permettent pas, pour l'heure, de faire apparaître un consensus sur la solution de correction du déséquilibre de classe à appliquer pour éviter un biais de prédiction. En effet, l'hétérogénéité des jeux de données, des méthodes et surtout l'absence d'optimisation des hyperparamètres des algorithmes utilisés ne permettent pas de tirer des conclusions fiables.

Une nouvelle problématique « Machine Learning » de recherche s'impose pour contribuer à l'émergence d'une nouvelle solution dans le cadre d'un déséquilibre de classe.

L'approche empirique de ce mémoire propose de vérifier les hypothèses « Machine Learning » suivantes :

- Le taux de sous-estimation peut être une mesure efficace de performance d'un algorithme de Machine Learning. Le choix de la mesure de la performance dépend de l'objectif du modèle. En effet, l'objectif peut être d'éviter de mal classer les observations positives. Cette hypothèse apporte une solution alternative aux mesures de performances habituelles.
- Le seuil de score de classification permet de définir les observations de classe d'intérêt positive. L'ajustement du seuil optimum pourrait-il se baser sur une métrique spécifique aux observations positives mal classées, i. e. le taux de sous-estimation ? Dans ce cas, l'ajustement a pour objectif de limiter le taux de faux négatifs.
- Le ratio de déséquilibre tolérable suit habituellement un ordre de grandeur de 5%-95%. Seule, une approche empirique de recherche de convergence du modèle

permet de préciser ce ratio. Pour cela, il convient de choisir la métrique de l'AUC PR pour mesurer les performances du modèle.

- La solution de correction du déséquilibre de classe peut être multiple. Le choix d'une méthode cost-sensitive peut sembler gourmande en ressources par rapport à une méthode de rééchantillonnage. Une solution consiste à combiner l'oversampling et la mise en œuvre du paramétrage intégré de XGBoost `scale_pos_weight`. L'approche empirique de ce mémoire apportera une réponse à la question de l'opportunité de la correction du déséquilibre de classe.

En conclusion, il convient d'être prudent et de maîtriser les algorithmes mis en œuvre, leur optimisation et le prétraitement des données nécessaires.

3.3 Prédiction des défaillances avec le projet Signaux Faibles

Ce chapitre présente succinctement le projet « Signaux Faibles » (République Française, 2021).

Ce projet porté par Bercy et ses partenaires, exploite les données administratives des partenaires du projet pour produire un outil d'analyse prédictive des difficultés des entreprises.

Les données utilisées proviennent de plusieurs sources (Camilleri, Joli, Viers, Coufour, & Ninucci, 2021) :

- Données Sirene Raison sociale, adresse, code APE, date de création etc.\
- Données DIRECCTE Autorisations et consommations d'activité partielle, recours à l'intérim, déclaration des mouvements de main-d'oeuvre
- Données URSSAF Données de défaillance, montant des cotisations, montant des dettes (part patronale, part ouvrière), demandes de délais de paiement, demandes préalables à l'embauche
- Données Banque de France 6 ratios financiers
- Données Diane Bilans et comptes de résultats. Permet d'enrichir les données financières de la Banque de France
- Données Altarès Base "paydex" sur les retards de paiements

Il s'agit de données importées telles que :

- Des données financières de la Banque de France,
- Des données financières issues des bilans déposés au greffe des tribunaux de commerce,
 - Structure et liquidité
 - Gestion
 - Productivité et rentabilité
 - Marge et valeur ajoutée
 - Compte de résultat
- Données sur l'activité partielle (Covid 19...)
 - Données de demandes d'activité partielle
 - Table des motifs de recours à l'activité partielle

- Table des périmètres du chômage
 - Table des recours antérieurs au chômage
 - Table des avis du CE
- Données de consommations d'activité partielle
- Table de correspondance entre compte administratif URSSAF et siret
- Données sur l'effectif
- Données sur les cotisations sociales et les débits
 - Fichier sur les cotisations
 - Fichiers sur les débits
 - Codes état du compte
 - Codes procédure collective
 - Codes opération historique
 - Code motif de l'écart négatif
- Données sur les délais
- Données sur les procédures collectives
- Données sur les CCSF
- Ellisphere (ellisphere, s.d.)
 - Lexique et explication des concepts clés sur les liens Ellisphere
 - Cas particuliers des Têtes de Groupe (TDG)
 - Structure du fichier
- Retards de paiements fournisseur de la base « Paydex » d'Altares

Plus précisément, le produit « Signaux Faibles » prédit un risque d'entrée en procédure collective des entreprises actuelles, à partir de leurs données passées et des trajectoires des entreprises ayant connu une défaillance. Il s'appuie sur un modèle récent de Machine Learning, pour fournir une prédiction de défaillance à 18 mois. Une liste des entreprises détectées en difficulté est produite.

Le code source est libre et fait l'objet d'un dépôt sur GitHub (GitHub / Signaux Faibles, 2021). La liste des données utilisées par le produit a inspiré la liste des variables exploitées dans la présente approche empirique.

4 Approche empirique de la prédiction de défaillance des entreprises

4.1 Méthodologie

L'approche empirique a suivi une méthodologie rigoureuse comportant les phases suivantes : définition de l'échantillon, recueil des données, visualisation des données, analyse brute des données, pré-traitement des données, analyse fonctionnelle des données constitution du dataset d'entraînement et de test et conception du modèle de Machine Learning.

4.1.1 Définition de l'échantillon des données des entreprises

La constitution de l'échantillon s'est faite en deux temps. Les Numéros de SIREN identifiant les entreprises ont été listés avant d'extraire l'échantillon en lui-même.

4.1.1.1 Choix du fournisseur des numéros de SIREN

La base spécifique SIREN a été choisie pour extraire une liste d'entreprises de type PME, ETI de préférence (data.gouv.fr).

4.1.1.2 Mode de sélection des entreprises actives et inactives du secteur industriel

Une sélection de 3000 Entreprises ciblées sur les codes division NAF 20 à 30 du secteur industriel a été conçue :

Code NAF	Intitulé de l'activité
20	Industries Chimiques
24	Métallurgie
25	Fabrication de produits métalliques sauf des machines et équipement
26	Fabrication de produits informatiques électroniques optiques
27	Fabrication de produits d'équipements électriques
28	Fabrication de machines et équipement n.c.a.
29	Industrie automobile
30	Fabrication d'autres matériels de transport

Les entreprises comportant plusieurs établissements ont été écartées du périmètre du projet, le projet ayant pour cible principale les PME du secteur industriel.

Des entreprises actives et des entreprises ayant cessé leur activité sont incluses dans la liste des numéros de SIREN.

4.1.1.3 Choix du fournisseur de données comptables et financières

La base Diane® (Bureau van Dijk) a été choisie pour fournir les données comptables et financières. Mais, les données comptables sont issues de la publication des comptes par les entreprises auprès des tribunaux de commerce. Ces dernières sont aussi disponibles sur le site de l'INPI (INPI, 2017).

Néanmoins, les entreprises ont la possibilité de délivrer leurs comptes sous couvert de confidentialité. Les données sont alors indisponibles dans la base Diane, comme sur le site de l'INPI.

Dans les faits, la ou les dernières années d'activité, en cas de cessation d'activité, ne font guère l'objet d'enregistrement de données comptables dans la base Diane.

4.1.1.4 Sélection des variables explicatives

La sélection des variables explicatives de la défaillance a été réalisée a priori.

Néanmoins, les variables intégrées dans les techniques d'analyse financière habituelles ont été sélectionnées de façon privilégiée. Ce sont les variables dites financières.

D'autres variables supposées impliquées dans le risque de défaillance ont été ajoutées : âge, délai crédit fournisseur, code NAF, code Banque, Ces sont les variables dites non financières. Dans l'ensemble, elles ont fait l'objet d'un pré-traitement décrit plus bas.

D'autres informations ont été requises pour les besoins de vérification de l'algorithme (Présence dans Diane, Nombre d'ES, Dernière année disponible, Nombre d'années, Type de compte, confidentialité -Date de clôture, ...) ou des besoins de prédictions futures. Elles ne sont pas décrites ici.

4.1.1.4.1 Variables financières

Les variables financières⁷ suivantes ont été exploitées :

- "Chiffre d'affaires\nkEUR\nDernière année disp.",
- "Chiffre d'affaires\nkEUR\nAnnée - 1",
- "Chiffre d'affaires\nkEUR\nAnnée - 2",
- "Chiffre d'affaires\nkEUR\nAnnée - 3",
- 'Valeur ajoutée\nkEUR\nDernière année disp.',
- 'Valeur ajoutée\nkEUR\nAnnée - 1',
- 'Valeur ajoutée\nkEUR\nAnnée - 2',
- 'Valeur ajoutée\nkEUR\nAnnée - 3',
- 'Bénéfice ou perte\nkEUR\nDernière année disp.',
- 'Bénéfice ou perte\nkEUR\nAnnée - 1',
- 'Bénéfice ou perte\nkEUR\nAnnée - 2',
- 'Bénéfice ou perte\nkEUR\nAnnée - 3',
- "Capacité d'autofinancement avant répartition\nkEUR\nDernière année disp.",
- "Capacité d'autofinancement avant répartition\nkEUR\nAnnée - 1",
- "Capacité d'autofinancement avant répartition\nkEUR\nAnnée - 2",
- "Capacité d'autofinancement avant répartition\nkEUR\nAnnée - 3",
- 'Fonds propres\nkEUR\nDernière année disp.',

⁷ Voir la définition des variables dans la documentation de Diane

- 'Fonds propres\nkEUR\nAnnée - 1',
- 'Fonds propres\nkEUR\nAnnée - 2',
- 'Fonds propres\nkEUR\nAnnée - 3',
- 'Capital social ou individuel\nkEUR\nDernière année disp.',
- 'Capital social ou individuel\nkEUR\nAnnée - 1',
- 'Capital social ou individuel\nkEUR\nAnnée - 2',
- 'Capital social ou individuel\nkEUR\nAnnée - 3',
- 'Fonds de roulement net global\nkEUR\nDernière année disp.',
- 'Fonds de roulement net global\nkEUR\nAnnée - 1',
- 'Fonds de roulement net global\nkEUR\nAnnée - 2',
- 'Fonds de roulement net global\nkEUR\nAnnée - 3',
- 'Endettement\n%\nDernière année disp.',
- 'Endettement\n%\nAnnée - 1',
- 'Endettement\n%\nAnnée - 2',
- 'Endettement\n%\nAnnée - 3',
- 'Liquidité réduite\n%\nDernière année disp.',
- 'Liquidité réduite\n%\nAnnée - 1',
- 'Liquidité réduite\n%\nAnnée - 2',
- 'Liquidité réduite\n%\nAnnée - 3',
- 'Rentabilité nette \n%\nDernière année disp.',
- 'Rentabilité nette \n%\nAnnée - 1',
- 'Rentabilité nette \n%\nAnnée - 2',
- 'Rentabilité nette \n%\nAnnée - 3',
- 'Rendement des capitaux propres nets\n%\nDernière année disp.',
- 'Rendement des capitaux propres nets\n%\nAnnée - 1',
- 'Rendement des capitaux propres nets\n%\nAnnée - 2',
- 'Rendement des capitaux propres nets\n%\nAnnée - 3',
- 'Total dettes : à 1 an au plus\nkEUR\nDernière année disp.',
- 'Total dettes : à 1 an au plus\nkEUR\nAnnée - 1',
- 'Total dettes : à 1 an au plus\nkEUR\nAnnée - 2',
- 'Total dettes : à 1 an au plus\nkEUR\nAnnée - 3',
- "Total dettes : à plus d'1 an et 5 ans au plus\nkEUR\nDernière année disp.",
- "Total dettes : à plus d'1 an et 5 ans au plus\nkEUR\nAnnée - 1",
- "Total dettes : à plus d'1 an et 5 ans au plus\nkEUR\nAnnée - 2",
- "Total dettes : à plus d'1 an et 5 ans au plus\nkEUR\nAnnée - 3",
- 'Total dettes : à plus de 5 ans\nkEUR\nDernière année disp.',
- 'Total dettes : à plus de 5 ans\nkEUR\nAnnée - 1',
- 'Total dettes : à plus de 5 ans\nkEUR\nAnnée - 2',
- 'Total dettes : à plus de 5 ans\nkEUR\nAnnée - 3',
- "TOTAL DE L'ACTIF\nkEUR\nDernière année disp.",
- "TOTAL DE L'ACTIF\nkEUR\nAnnée - 1",
- "TOTAL DE L'ACTIF\nkEUR\nAnnée - 2",

- "TOTAL DE L'ACTIF\nekEUR\nAnnée - 3",
- "Excédent brut d'exploitation\nekEUR\nDernière année disp.",
- "Excédent brut d'exploitation\nekEUR\nAnnée - 1",
- "Excédent brut d'exploitation\nekEUR\nAnnée - 2",
- "Excédent brut d'exploitation\nekEUR\nAnnée - 3",
- 'Besoins (res.) en fonds de roul.\nekEUR\nDernière année disp.',
- 'Besoins (res.) en fonds de roul.\nekEUR\nAnnée - 1',
- 'Besoins (res.) en fonds de roul.\nekEUR\nAnnée - 2',
- 'Besoins (res.) en fonds de roul.\nekEUR\nAnnée - 3',
- 'Equilibre financier\nDernière année disp.',
- 'Equilibre financier\nAnnée - 1',
- 'Equilibre financier\nAnnée - 2',
- 'Equilibre financier\nAnnée - 3',
- 'Indépendance fin.\n%\nDernière année disp.',
- 'Indépendance fin.\n%\nAnnée - 1',
- 'Indépendance fin.\n%\nAnnée - 2',
- 'Indépendance fin.\n%\nAnnée - 3',
- "Taux d'endettement\n%\nDernière année disp.",
- "Taux d'endettement\n%\nAnnée - 1",
- "Taux d'endettement\n%\nAnnée - 2",
- "Taux d'endettement\n%\nAnnée - 3",
- 'Capacité de remboursement\nDernière année disp.',
- 'Capacité de remboursement\nAnnée - 1',
- 'Capacité de remboursement\nAnnée - 2',
- 'Capacité de remboursement\nAnnée - 3',
- "Capacité d'autofin.\n%\nDernière année disp.",
- "Capacité d'autofin.\n%\nAnnée - 1",
- "Capacité d'autofin.\n%\nAnnée - 2",
- "Capacité d'autofin.\n%\nAnnée - 3",
- 'Rentabilité économique\n%\nDernière année disp.',
- 'Rentabilité économique\n%\nAnnée - 1',
- 'Rentabilité économique\n%\nAnnée - 2',
- 'Rentabilité économique\n%\nAnnée - 3',
- 'Taux de marge commerciale\n%\nDernière année disp.',
- 'Taux de marge commerciale\n%\nAnnée - 1',
- 'Taux de marge commerciale\n%\nAnnée - 2',
- 'Taux de marge commerciale\n%\nAnnée - 3',
- 'Taux de valeur ajoutée\n%\nDernière année disp.',
- 'Taux de valeur ajoutée\n%\nAnnée - 1',
- 'Taux de valeur ajoutée\n%\nAnnée - 2',
- 'Taux de valeur ajoutée\n%\nAnnée - 3',
- "Poids des BFR d'exploitation\n%\nDernière année disp.",

- "Poids des BFR d'exploitation\n%\nAnnée - 1",
- "Poids des BFR d'exploitation\n%\nAnnée - 2",
- "Poids des BFR d'exploitation\n%\nAnnée - 3",
- #"Variation du besoin en fonds de roulement liée à l'activité\nkEUR\nDernière année disp.",
- #"Variation du besoin en fonds de roulement liée à l'activité\nkEUR\nAnnée - 1",
- #"Variation du besoin en fonds de roulement liée à l'activité\nkEUR\nAnnée - 2",
- #"Variation du besoin en fonds de roulement liée à l'activité\nkEUR\nAnnée - 3".

4.1.1.4.2 Variables non financières

Les variables non financières suivantes ont été exploitées telles quelles par le modèle :

- 'NAF Rév. 2, code principal (code)',
- "Nombre d'employés\nDernière année disp.",
- "Nombre d'employés\nAnnée - 1",
- "Nombre d'employés\nAnnée - 2",
- "Nombre d'employés\nAnnée - 3",
- 'Crédit clients\njours\nDernière année disp.',
- 'Crédit clients\njours\nAnnée - 1',
- 'Crédit clients\njours\nAnnée - 2',
- 'Crédit clients\njours\nAnnée - 3',
- 'Crédit fournisseurs\njours\nDernière année disp.',
- 'Crédit fournisseurs\njours\nAnnée - 1',
- 'Crédit fournisseurs\njours\nAnnée - 2',
- 'Crédit fournisseurs\njours\nAnnée - 3',

D'autres variables non financières ont fait l'objet d'un pré-traitement avant d'être exploitées :

- #'Date de création',
- #'Date associée à la situation juridique',
- #'NAF Rév. 2, code principal (code)',
- #'Code de la banque',
- #'Tranche d'effectif salarié de l'entreprise',
- #'Statut juridique'.

La variable non financière « Statut juridique » est la variable à expliquer.

La variable non financière « Procédure collective » n'a pas été retenue comme variable à expliquer. Néanmoins, elle a fait l'objet d'une analyse.

4.1.1.4.3 Profondeur historique requise

Une profondeur historique de 4 ans a été retenue. Les données historisées apparaissent sous forme de duplication des variables choisies sur les années N, N-1, N-2 et N-3.

La profondeur historique est relative à l'année de publication des comptes la plus récente.

Dans le cas de cessation d'activité, comme dans le cas d'activité en cours, les données requises concernent les quatre dernières années de publication des comptes.

4.1.2 Recueil des données

4.1.2.1 Téléchargement des données de SIREN

L'extraction des numéros de SIREN a été réalisée à partir de la base SIREN du 01/06/2021.

La base des unités légales, unité légale signifiant la plus petite entité juridique de l'entreprise, a été utilisée.

La liste obtenue a, par la suite, été filtrée deux fois.

Le premier filtre a permis de cibler le type d'entreprise PME et ETI.

Le second filtre a ciblé les codes de division d'activité NAF d'intérêt pour le secteur industriel et retenus lors de la phase précédente de définition de l'échantillon.

4.1.2.2 Oversampling à des fins de rééquilibrage des classes de défaillances

Le déséquilibre de classe limite les performances des algorithmes de classification. Un oversampling est effectué pour résoudre cela. Un rééquilibrage au profit des entreprises ayant cessé leur activité permet d'atteindre une proportion suffisante d'exemples de défaillance.

Pour cela, des entreprises ayant cessé leur activité sont injectées dans la liste initiale des entreprises pour constituer une liste de référence de numéros de SIREN.

Le POC a porté sur un extrait seulement de cette liste de référence. Cet extrait a été limité à 3000 observations.

Après oversampling (et extraction finale sur la base Diane), la proportion d'entreprises ayant cessé leur activité est la suivante :

Ratio de déséquilibre = $374 / 2348 = 15,93 \%$

C'est le ratio de déséquilibre du dataset exploité.

4.1.2.3 Extraction et téléchargement des données comptables et financières

L'extraction des données financières et comptables et le téléchargement des fichiers de données ont été effectués, depuis la base Diane, à partir de trois listes de 1000 SIREN d'entreprises chacun.

Les opérations d'extraction ont eu lieu à différentes reprises du 16/06/2021 au 31/08/2021.

4.1.2.4 Lecture des fichiers de données

Différentes normalisations de l'information « données manquantes » sont apparues, lors d'une première visualisation, dans les fichiers de données extraits. Les valeurs suivantes sont retrouvées : NaN, blanc, « n.d. », « n.s. », « n.d. », ou *diverses indications d'absence de données*.

Elles sont toutes converties en une unique normalisation : NaN. Une option de lecture du fichier excel (na_values = missing_values) a été utilisée à cette fin.

En effet, l'algorithme XGBoost sait traiter l'absence des données dont l'indication est notée suivant cette norme.

Il est à noter que les valeurs renseignées à 0 ont été considérées significatives et conservées en l'état.

4.1.3 Visualisation et présentation des données du dataset

Une première visualisation des données a été faite. Elle a permis plusieurs constatations :

- Le volume insuffisant de procédures collectives dans le dataset,
- La proportion d'entreprises inactives à savoir 374 entreprises inactives sur un total de 2 348 entreprises

4.1.4 Analyse brute des données

4.1.4.1 Qualité des données

La valorisation des données est normalisée pour chaque variable.

Les données financières sont fournies en KEuros.

4.1.4.2 Données vides

```
print ("taux de valeurs nulles dans l'export de Diane : ", nb_nan/(df_features.shape[0]*df_features.shape[1]))
taux de valeurs nulles dans l'export de Diane :  0.15831638677512463
```

Le ratio de valeurs manquantes dans le dataset est de 15,83%.

Nombre d'entreprises comportant une valeur manquante par variable dataset :

une absence de publication de compte dans Diane de l'ordre de 17,70 %
(177 SIREN manquants sur 1000)

nombre de lignes SIREN comportant des variables vides (NaN Not a Number) :

Variation du besoin en fonds de roulement liée à
l'activité\нкEUR\нAnnée - 3 2348

Date de la fin de procédure de sauvegarde
2348

Variation du besoin en fonds de roulement liée à
l'activité\нкEUR\нAnnée - 1 2348

Variation du besoin en fonds de roulement liée à
l'activité\нкEUR\нDernière année disp. 2348

Variation du besoin en fonds de roulement liée à
l'activité\нкEUR\нAnnée - 2 2348

Capitalisation boursière\нкEUR
2336

Fin de la procédure collective
2224

Début de la procédure collective
2105

Date associée à la situation juridique
1980

Situation juridique
 1978
 Taux de marge commerciale\n%\nDernière année disp.
 1727
 Taux de marge commerciale\n%\nAnnée - 2
 1680
 Taux de marge commerciale\n%\nAnnée - 1
 1675
 Taux de marge commerciale\n%\nAnnée - 3
 1656
 Nombre d'employés\nDernière année disp.
 1540
 Nombre d'employés\nAnnée - 1
 1413
 Nombre d'employés\nAnnée - 2
 1339
 Confidentialité - Date de clôture
 1307
 Nombre d'employés\nAnnée - 3
 1224
 Rendement des capitaux propres nets\n%\nDernière année disp.
 1056
 Rentabilité nette \%\nDernière année disp.
 988
 Bénéfice ou perte\nkEUR\nDernière année disp.
 964
 Rendement des capitaux propres nets\n%\nAnnée - 1
 800
 Rentabilité nette \%\nAnnée - 1
 747
 Bénéfice ou perte\nkEUR\nAnnée - 1
 726
 Description de l'activité
 712
 Capacité d'autofin.\n%\nDernière année disp.
 586
 Rendement des capitaux propres nets\n%\nAnnée - 2
 580
 Taux de valeur ajoutée\n%\nDernière année disp.
 559
 Rentabilité économique\n%\nDernière année disp.
 559
 Poids des BFR d'exploitation\n%\nDernière année disp.
 558
 Endettement\n%\nDernière année disp.
 538
 Code de la banque
 537
 Nom de la banque
 537
 Crédit clients\njours\nDernière année disp.
 536
 Capacité de remboursement\nDernière année disp.
 534
 Liquidité réduite\n%\nDernière année disp.
 516
 Rentabilité nette \%\nAnnée - 2
 503

Crédit fournisseurs\njours\nDernière année disp.
 502
 Capacité d'autofin.\n%\nAnnée - 1
 501
 Valeur ajoutée\nkEUR\nDernière année disp.
 489
 Fonds de roulement net global\nkEUR\nDernière année disp.
 489
 Capacité d'autofinancement avant répartition\nkEUR\nDernière année
 disp. 489
 Bénéfice ou perte\nkEUR\nAnnée - 2
 488
 Rentabilité économique\n%\nAnnée - 1
 486
 Taux de valeur ajoutée\n%\nAnnée - 1
 486
 Poids des BFR d'exploitation\n%\nAnnée - 1
 485
 Capital social ou individuel\nkEUR\nDernière année disp.
 484
 Fonds propres\nkEUR\nDernière année disp.
 484
 Chiffre d'affaires\nkEUR\nDernière année disp.
 481
 Crédit clients\njours\nAnnée - 1
 471
 Capacité de remboursement\nAnnée - 1
 471
 Chiffre d'affaires\nkEUR\nAnnée - 1
 453
 Crédit fournisseurs\njours\nAnnée - 1
 441
 Capacité d'autofin.\n%\nAnnée - 2
 418
 Rendement des capitaux propres nets\n%\nAnnée - 3
 409
 Taux de valeur ajoutée\n%\nAnnée - 2
 398
 Poids des BFR d'exploitation\n%\nAnnée - 2
 398
 Rentabilité économique\n%\nAnnée - 2
 398
 Crédit clients\njours\nAnnée - 2
 391
 Capacité de remboursement\nAnnée - 2
 388
 Chiffre d'affaires\nkEUR\nAnnée - 2
 374
 Crédit fournisseurs\njours\nAnnée - 2
 368
 Capacité d'autofin.\n%\nAnnée - 3
 361
 Rentabilité économique\n%\nAnnée - 3
 347
 Taux de valeur ajoutée\n%\nAnnée - 3
 347
 Poids des BFR d'exploitation\n%\nAnnée - 3
 347

Crédit clients\njours\nAnnée - 3
 347
 Capacité de remboursement\nAnnée - 3
 336
 Rentabilité nette \n%\nAnnée - 3
 335
 Bénéfice ou perte\nkEUR\nAnnée - 3
 320
 Endettement\n%\nAnnée - 1
 320
 Chiffre d'affaires\nkEUR\nAnnée - 3
 318
 Crédit fournisseurs\njours\nAnnée - 3
 317
 Liquidité réduite\n%\nAnnée - 1
 304
 Capacité d'autofinancement avant répartition\nkEUR\nAnnée - 1
 285
 Fonds de roulement net global\nkEUR\nAnnée - 1
 285
 Valeur ajoutée\nkEUR\nAnnée - 1
 285
 Fonds propres\nkEUR\nAnnée - 1
 282
 Capital social ou individuel\nkEUR\nAnnée - 1
 282
 Endettement\n%\nAnnée - 2
 174
 Liquidité réduite\n%\nAnnée - 2
 149
 Fonds de roulement net global\nkEUR\nAnnée - 2
 134
 Capacité d'autofinancement avant répartition\nkEUR\nAnnée - 2
 134
 Valeur ajoutée\nkEUR\nAnnée - 2
 134
 Capital social ou individuel\nkEUR\nAnnée - 2
 131
 Total dettes : à plus de 5 ans\nkEUR\nAnnée - 3
 131
 Total dettes : à plus d'1 an et 5 ans au plus\nkEUR\nAnnée - 3
 130
 Total dettes : à 1 an au plus\nkEUR\nAnnée - 3
 130
 Fonds propres\nkEUR\nAnnée - 2
 129
 Total dettes : à plus d'1 an et 5 ans au plus\nkEUR\nAnnée - 2
 121
 Total dettes : à 1 an au plus\nkEUR\nAnnée - 2
 120
 Total dettes : à plus de 5 ans\nkEUR\nAnnée - 2
 120
 Indépendance fin.\n%\nAnnée - 3
 113
 Equilibre financier\nAnnée - 3
 108
 Total dettes : à plus de 5 ans\nkEUR\nAnnée - 1
 102

Total dettes : à 1 an au plus\nkEUR\nAnnée - 1
 101
 Total dettes : à plus d'1 an et 5 ans au plus\nkEUR\nAnnée - 1
 101
 Indépendance fin.\n%\nAnnée - 2
 91
 Equilibre financier\nAnnée - 2
 91
 Endettement\n%\nAnnée - 3
 77
 Total dettes : à 1 an au plus\nkEUR\nDernière année disp.
 75
 Indépendance fin.\n%\nAnnée - 1
 75
 Total dettes : à plus d'1 an et 5 ans au plus\nkEUR\nDernière année
 disp. 74
 Total dettes : à plus de 5 ans\nkEUR\nDernière année disp.
 74
 Equilibre financier\nAnnée - 1
 73
 Taux d'endettement\n%\nAnnée - 3
 66
 Besoins (res.) en fonds de roul.\nkEUR\nAnnée - 3
 65
 TOTAL DE L'ACTIF\nkEUR\nAnnée - 3
 65
 Excédent brut d'exploitation\nkEUR\nAnnée - 3
 65
 Indépendance fin.\n%\nDernière année disp.
 62
 Equilibre financier\nDernière année disp.
 62
 Liquidité réduite\n%\nAnnée - 3
 61
 Taux d'endettement\n%\nAnnée - 2
 49
 Besoins (res.) en fonds de roul.\nkEUR\nAnnée - 2
 47
 TOTAL DE L'ACTIF\nkEUR\nAnnée - 2
 47
 Excédent brut d'exploitation\nkEUR\nAnnée - 2
 47
 Valeur ajoutée\nkEUR\nAnnée - 3
 37
 Fonds de roulement net global\nkEUR\nAnnée - 3
 37
 Capacité d'autofinancement avant répartition\nkEUR\nAnnée - 3
 37
 Fonds propres\nkEUR\nAnnée - 3
 35
 Capital social ou individuel\nkEUR\nAnnée - 3
 35
 Taux d'endettement\n%\nAnnée - 1
 26
 Besoins (res.) en fonds de roul.\nkEUR\nAnnée - 1
 24
 Excédent brut d'exploitation\nkEUR\nAnnée - 1
 24

TOTAL DE L'ACTIF\nkEUR\nAnnée - 1
 24
 Capital social courant\nkEUR
 2
 Taux d'endettement\n%\nDernière année disp.
 1
 Numéro Siren
 0
 Tranche d'effectif salarié de l'entreprise
 0
 Procédure collective
 0
 Raison sociale
 0
 Statut juridique
 0
 Numéro Siret
 0
 Département
 0
 Taille de l'unité urbaine (code)
 0
 Région
 0
 Pays
 0
 Code postal
 0
 Ville
 0
 Score AFDCC\nAnnée - 2
 0
 Score AFDCC\nAnnée - 3
 0
 Nombre d'années
 0
 Score AFDCC\nAnnée - 1
 0
 Score AFDCC\nDernière année disp.
 0
 Date de création
 0
 Cotée/Décotée/Non cotée
 0
 Présence dans Diane
 0
 Nombre d'ES
 0
 NAF Rév. 2, code principal (code)
 0
 NAF Rév. 2, code principal
 0
 Dernière année disponible
 0
 Besoins (res.) en fonds de roul.\nkEUR\nDernière année disp.
 0
 Excédent brut d'exploitation\nkEUR\nDernière année disp.
 0

```
TOTAL DE L'ACTIF\nkEUR\nDernière année disp.  
0  
Type de comptes  
0  
Nom de l'entreprise  
0
```

Nombre d'entreprise par statut juridique

```
df_features["Statut juridique"].value_counts()  
  
Actif      1974  
Inactif    374  
Name: Statut juridique, dtype: int64
```

4.1.5 Pré-traitement des données

La méthodologie suivie comprend une phase de pré-traitement des données pour amélioration de la convergence du modèle. Le pré-traitement a différents objectifs :

- Calcul de l'âge de l'entreprise, à la date du jour où à la date de cessation d'activité
 - Calcul de l'horizon de défaillance défini par le délai écoulé entre les derniers comptes publiés et la date associée aux différentes situations juridiques de cessation d'activité
 - Calcul intermédiaire de la dernière année disponible en nombre de jours
 - Extraction de l'année de défaillance
- Calcul des tendances des informations pertinentes : calcul des tendances année n-3, n-2, n-1 lorsque les données étaient disponibles pour les deux années concernées.
- Calcul de ratios importants d'analyse financière : calcul de ratios supplémentaires pour a-dimensionnaliser certaines informations et les rendre comparables d'une entreprise à une autre.
 - A-dimensionnement par rapport au Chiffre d'affaires pour les principales variables financières.
 - A-dimensionnement du Chiffre d'affaires par rapport à l'effectif lorsque la donnée était disponible.
 - Encodage des variables catégorielle : code NAF, code Banque, Tranche d'effectif salarié de l'entreprise, statut juridique

Il est à noter que le pré-traitement a peu porté sur l'absence des données. Les choix suivants ont été faits :

- Conservation des lignes et colonnes comportant partiellement des NaN
- Exclusion de deux colonnes de variables comportant 100% de données manquantes : Score AFDCC et Variation du besoin en fonds de roulement liée à l'activité.

Remarques : Les variables supplémentaires créées n'ont pas fait l'objet de tri ou sélection mais ont été ajoutées aux précédentes.

Deux variables ont été utilisées, puis exclues du modèle de prédiction pour cause de biais :

- #'Dernière année disponible',

- #"Nombre d'années".

4.1.6 Analyse fonctionnelle des données

4.1.6.1 Recherche d'une procédure collective

On constate une distorsion du nombre de procédures collectives 47 et du nombre des entreprises inactives 374 sur le dataset de 2 348 entreprises. C'est une limite du dataset découverte en qualification des données.

Dans celui-ci, le statut juridique Inactif correspond toujours, hormis une seule entreprise (est-ce un cas d'erreur de renseignement ?) à une situation juridique renseignée et vice-versa.

Les différentes situations juridiques enregistrées sont les suivantes : cessation d'activité, dissolution, dissolution anticipée, entreprise absorbée, liquidation amiable, liquidation judiciaire, liquidation judiciaire simplifiée, plan de cession, transmission universelle du patrimoine. Malheureusement, les sources d'informations et de définition de ces situations sont insuffisantes. On peut noter, néanmoins, que la cessation d'activité, la dissolution, la dissolution anticipée, la liquidation amiable, la liquidation judiciaire et la liquidation judiciaire simplifiée ont trait à la fermeture d'une entreprise à l'amiable ou non.

Nombre de Numéro Siren		
Statut juridique	Situation juridique	Total
Actif	(vide)	714
Total Actif		714
Inactif	Cessation d'activité	14
	Dissolution	5
	Dissolution anticipée	3
	Entreprise absorbée	3
	Liquidation amiable	9
	Liquidation judiciaire	62
	Liquidation judiciaire simplifiée	6
	Plan de cession	1
	Transmission universelle du patrimoine	5
	(vide)	1
Total Inactif		109
(vide)	(vide)	
Total (vide)		
Total général		823

Figure 14 Le nombre d'entreprises par activité et situation juridique

Hormis les cas de transmissions de patrimoine et les cas d'entreprises absorbées, peu nombreux dans le dataset, toutes les autres situations juridiques ont été jugées à risque concernant la continuité de la production industrielle de ces entreprises.

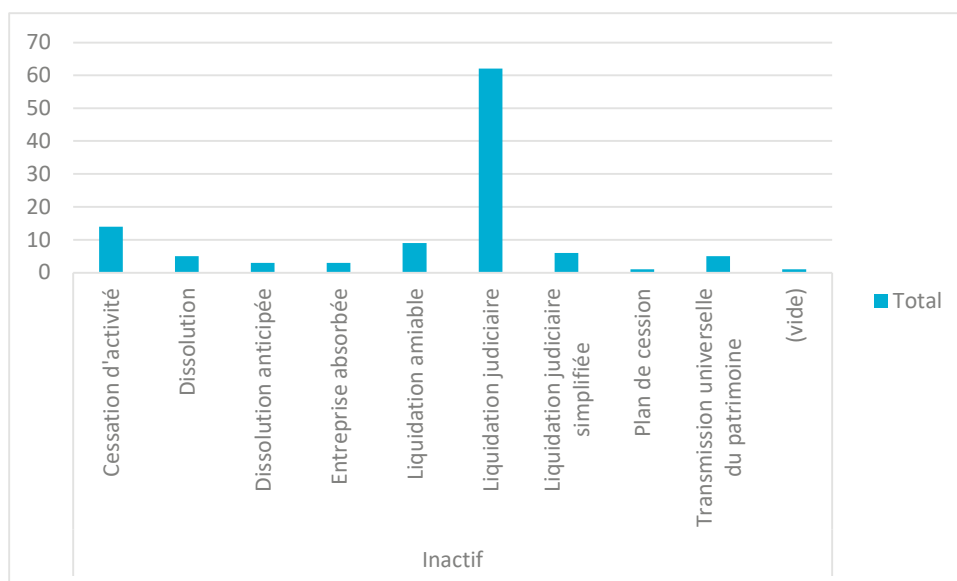


Figure 15 Nombre d'entreprises inactives par situation juridique

Dans un contexte de prévision du maintien opérationnel des équipements de la DGA, c'est ainsi que la prévision du maintien du statut juridique « actif » de l'entreprise, ou son corollaire le risque de survenance du statut juridique « inactif », a été privilégiée.

Par ailleurs, très peu d'entrées en procédure collective sont dénombrées et anticipées durant la période d'activité de l'entreprise (2 pour 823 Siren), et relativement peu sont enregistrées en cas d'inactivité (36 pour 109 Siren).

On pourrait en déduire que l'entrée en procédure collective n'est pas l'unique cause de la cessation d'activité.

Statut juridique		Procédure collective	Total
Actif	Non		712
	Oui		2
Total Actif			714
Inactif	Non		73
	Oui		36
Total Inactif			109
(vide)	(vide)		
Total (vide)			
Total général			823

Figure 16 Nombre d'entreprises en procédures collectives

En effet, les difficultés des entreprises hors état de cessation de paiement (mandat ad hoc et procédure de sauvegarde) ou de moins de 45 jours d'état de cessation de paiements (procédure de conciliation) ne sont pas identifiés et quantifiés dans ce dataset. En théorie, la confidentialité du mandat ad hoc et de la procédure de conciliation ne le permet pas. En particulier, les entreprises en état de cessation de paiement de moins de 45 jours et en cours de procédure de conciliation, sont invisibles sur cette visualisation. Or, la montée en puissance de ce type de procédure a été relevée précédemment dans la revue de littérature. Il en résulte une regrettable absence de données qualifiant ces situations dans le dataset.

Cependant, Il faut garder à l'esprit son effet protecteur de l'entreprise en difficulté. Un mandataire précise qu'une perte de confidentialité de sa situation délicate entraîne un risque de perte du crédit fournisseur⁸ qui représente en France (à la différence de nombreux pays européens), une très grande partie du crédit aux Entreprises (Guerin & Branchu-Bord, 2021).

En conclusion, seules, les différentes situations juridiques et le statut juridique général (actif versus inactif) sont quantifiées distinctement. Les procédures collectives ne sont malheureusement pas distinguées : les différentes procédures de sauvegarde (la sauvegarde, la sauvegarde accélérée, la sauvegarde financière accélérée), la procédure de redressement et les procédures de liquidation judiciaire (la procédure de liquidation judiciaire et la procédure de liquidation judiciaire simplifiée). Le mandat ad hoc, la procédure de conciliation et la procédure de rétablissement professionnelle ne sont pas, non plus, identifiées dans le dataset.

C'est pourquoi, la caractérisation de la cessation d'activité de l'entreprise a été retenue.

En conséquence, la caractérisation de l'entrée en procédure collective a été abandonnée.

4.1.6.2 Analyse de la variable ajoutée « horizon de défaillance »

L'horizon de défaillance est de 18 mois en moyenne. 75% des défaillances ont lieu dans les 2 ans qui suivent la dernière publication des comptes

Ces valeurs de l'horizon (moyenne et quartiles) donnent l'horizon du modèle de prédiction de défaillance des entreprises conçu dans cette approche empirique.

4.1.7 Constitution du dataset d'entraînement et de test

4.1.7.1 Sélection opérationnelle des variables cibles et explicatives

Abandon des données jugées moins pertinentes dans un premier temps (adresse, ...)

Abandon des données biaisées telles que la dernière année et le nombre d'années de présence des comptes dans la base Diane.

4.1.7.2 Train_test_split et paramètre stratify

```
X_train, X_test, y_train, y_test = train_test_split(dataset,
                                                    dataset[target],
                                                    test_size = 0.35,
                                                    stratify=dataset[target],
                                                    random_state=42 #Contrôle le brassage appliqué aux données avant d'appliquer
                                                    )
```

⁸ Le crédit fournisseur est un crédit accordé à l'acheteur par le fournisseur dans le cadre de son contrat commercial. Selon les pays, les branches d'activité et les accords particuliers passés entre un fournisseur et son client, il peut être convenu - en matière commerciale - que le fournisseur soit payé au terme d'un délai déterminé (30 à 90 jours le plus souvent, voire dans certains cas plus longtemps). (Wikipédia, s.d.)

4.1.7.3 Shape du dataset de formation et de test

```
print("shape du dataset de train: ", X_train.shape)
```

```
shape du dataset de train: (1457, 311)
```

```
print("shape du dataset de test: ", X_test.shape)
```

```
shape du dataset de test: (786, 311)
```

4.1.8 Conception du modèle de Machine Learning avec XGBoost

4.1.8.1 Hyperparamètres du modèle XGBoost

Une première série d'hyperparamètres est fixée à des valeurs connus dans la littérature

```
Xgboost = XGBClassifier(  
    objective= 'binary:logistic', # type de classifieur binaire logistique et fonction de cout logloss induite  
    nthread=4,  
    seed=42, # valeur de départ aléatoire fixée permet la reproductibilité de l'algo pour la documentation  
    scale_pos_weight=estimate, # valeur de gestion du déséquilibre de classe de défaillance  
    max_depth = 5, # valeur habituellement trouvée dans la littérature  
    learning_rate = 0.1, # valeur calculée par le gridsearch dans 2021_08_21_DIANE_EXTRACT2348_GESTION-Algo_uni  
    reg_lambda = 1000.0, # valeur calculée par le gridsearch dans 2021_08_21_DIANE_EXTRACT2348_GESTION-Algo_uni  
    n_estimators = 140 # valeur calculée par le gridsearch dans 2021_08_21_DIANE_EXTRACT2348_GESTION-Algo_uniqu  
)
```

4.1.8.2 Choix des autres hyperparamètres importants

Utilisation du GridSearchCV pour sélectionner les hyperparamètres donnant les meilleures performances

```
parameters = {  
    'subsample': np.arange(0.5, 1.0, 0.2), # ratio de sous-échantillon des instances d'entraînement en prévention de l'overfittin  
    'colsample_bytree': np.arange(0.4, 1.0, 0.2), # rapport de sous-échantillon de colonnes lors de la construction de chaque ar  
    'colsample_bylevel': np.arange(0.4, 1.0, 0.2), # rapport de sous-échantillon de colonnes pour chaque niveau  
    'colsample_bynode': np.arange(0.4, 1.0, 0.2) # ratio de sous-échantillon de colonnes pour chaque nœud (split)  
}
```

4.1.8.3 Grid search et recherche des meilleurs paramètres

4.1.8.3.1 Validation croisée

4.1.8.3.2 Paramètres du GridSearchCV importants

```
grid_search_Xgboost = GridSearchCV(  
    estimator=Xgboost,  
    param_grid=parameters,  
    scoring = 'recall',  
    n_jobs = None,  
    cv = 5,  
    return_train_score=True,  
    verbose=True  
)
```

4.1.8.4 Recherche des hyperparamètres avec `best_estimator_` du `GridSearchCV`

Utilisation des paramètres `best_estimator` et les autres paramètres

```
grid_search_Xgboost.best_estimator_  
  
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=0.4,  
              colsample_bynode=0.4, colsample_bytree=0.6000000000000001,  
              gamma=0, gpu_id=-1, importance_type='gain',  
              interaction_constraints='', learning_rate=0.1, max_delta_step=0,  
              max_depth=5, min_child_weight=1, missing=nan,  
              monotone_constraints='()', n_estimators=140, n_jobs=4, nthread=4,  
              num_parallel_tree=1, random_state=42, reg_alpha=0,  
              reg_lambda=1000.0, scale_pos_weight=5.872641509433962, seed=42,  
              subsample=0.7, tree_method='exact', validate_parameters=1,  
              verbosity=None)
```

4.1.8.5 Recherche du meilleur seuil de classification

D'abord, une prédiction est réalisée sur la base du seuil de classification par défaut. Puis deux autres prédictions sont réalisées, l'une maximisant la performance F-Mesure, l'autre minimisant le taux de sous-estimation.

4.1.8.5.1 Recherche du seuil optimum pour maximiser la F-Mesure

Après traçage de la courbe Précision – Recall, une recherche du seuil optimum pour maximiser la F-Mesure permet de le localiser sur cette courbe.

La recherche du seuil consiste à convertir les mesures Précision et Recall issues du modèle en F-Mesure d'après sa formule de calcul. Puis le seuil optimum est déterminé comme étant le seuil à l'origine du maximum de la F-Mesure.

Ensuite, les étiquettes de classe des observations peuvent être calculées en fonction du nouveau seuil.

4.1.8.5.2 Recherche du seuil optimum pour minimiser le taux de sous-estimation

Un seuil de classification minimisant le taux de sous-estimation a été fixée spécifiquement pour cette approche empirique. En effet, le taux de sous-estimation représente le taux d'entreprises réellement défaillantes non détectées en tant que telles. Or, la non-détection des défaillances avérées pourrait être le premier facteur de refus du modèle de prédiction. Il est plus coûteux de ne pas détecter que de sur-détecter les défaillances dans cette approche empirique.

La méthode de recherche du seuil optimum pour minimiser le taux de sous-estimation est décrite ci-après.

Après traçage de la courbe ROC, une recherche du seuil optimum minimisant le taux de sous-estimation permet de le localiser sur cette courbe.

Pour mémoire, cette courbe porte le True Positive Rate sur l'axe Y et le False Positive Rate sur l'axe X.

La recherche du seuil consiste à exploiter le True Positive Rate issu du modèle pour en extraire une nouvelle borne.

Un taux maximum de 5% de False Négatives a été fixée arbitrairement pour les besoins de cette approche empirique. Puis, une valeur maximum de TPR et un seuil optimum en sont déduits comme suit :

$$\text{FNR} < 0,05$$

$$\text{Or FNR} = 1 - \text{TPR}$$

$$\text{Donc TPR} > 0,95$$

Puis, le seuil optimum est déterminé comme étant le premier seuil à l'origine de cette borne minimum du TPR.

Remarque : Dans cette approche, la courbe ROC n'est pas exploitée pour calculer l'AUC. C'est la courbe Précision Rappel qui fournit la mesure de performance AUC PR adéquate. Néanmoins, la courbe ROC permet de visualiser le taux de sous-estimation maximum tolérée.

4.2 Résultats

4.2.1 Mesure des résultats en termes de performance

4.2.1.1 *Mesure des résultats « Machine Learning » en termes de performance : Utilisation de la mesure du taux de sous-estimation d'un modèle et méthode d'ajustement d'un seuil de classification en fonction du taux de sous-estimation*

4.2.1.1.1 Nombre de défaillances prédites en fonction des Bénéfices et Pertes

Le tableau croisé dynamique suivant croise les prédictions de défaillance avec la variable explicative la plus importante à savoir les Bénéfices et Pertes.

Défaillances prédites	Nombre de Numéro Siren	Total Bénéfice ou perte Dernière année disp. /CA en Ke
0	365	13,89594932
Actif	359	13,75503046
Inactif	6	0,140918858
1	421	-11,93142558
Actif	312	-1,116557899
Inactif	109	-10,81486768
(vide)		
(vide)		
Total général	786	1,964523734

Figure 17 Pertes et Bénéfices des entreprises prédites à risque de défaillance ou non

4.2.1.1.2 Les matrices de confusion résultantes en fonction des différents seuils de classification utilisés

Trois seuils sont comparés : le seuil de classification défini par défaut à la valeur 0,5, le seuil calculé pour optimiser la mesure F-Score et le seuil calculé pour minimiser le taux de sous-estimation.

Les 3 matrices de confusion résultantes sont les suivantes :

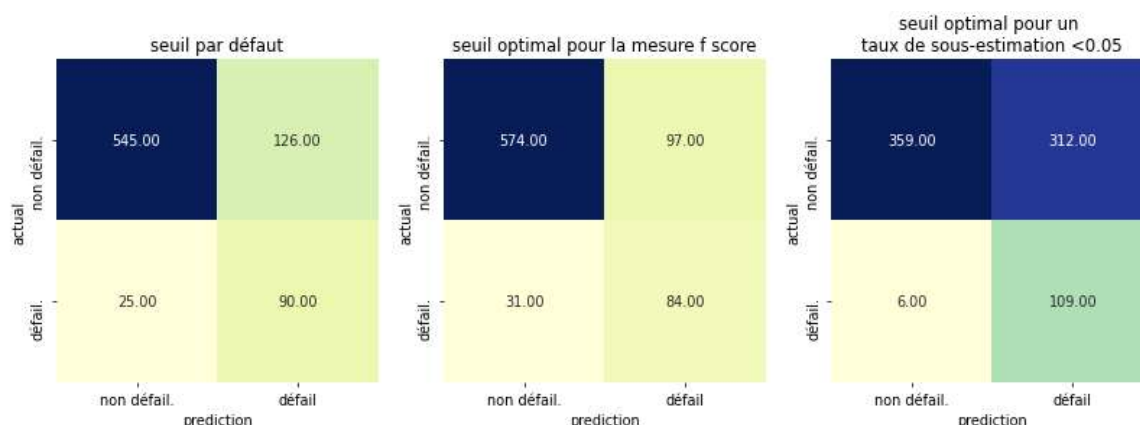


Figure 18 Supériorité du seuil de classification minimisant le taux de sous-estimation

La matrice de confusion de droite se lit de la façon suivante :

- Les prédictions confirmées par la réalité sont les suivantes :
 - Nombre d'entreprises non-défaillantes confirmées : 395
 - Nombre d'entreprises défaillantes confirmées 109
- Les prédictions erronées sont les suivantes :
 - Nombre d'entreprises défaillantes non détectées 6
- Les « vraies prédictions » sont les suivantes :
 - Nombre d'entreprises non-défaillantes à ce jour mais détectées à fort risque de défaillance en 2022⁹ : 276

4.2.1.1.3 Performances du modèle de prédiction des défaillances

Les performances obtenues sont les suivantes :

- Un seuil de classification des défaillances a été fixé afin d'obtenir un taux de sous-estimation inférieur à 5 %.
- Le modèle obtient un taux de sous-estimation de 0,04378 et une mesure de ROC AUC de 0,873609.

⁹ Prévision du risque pour la période du 01/01/2021 au 30/06/2022, durant laquelle 75% des défaillances prévues sont attendues

	model_name	bal. accuracy	tx de sous-estimation	ROC AUC
1	Xgboost seuil par défaut	0.796669	0.217391	0.873609
2	Xgboost seuil / f score	0.792937	0.269565	0.873609
3	Xgboost seuil / tx de sous-estimation	0.741424	0.043478	0.873609

Figure 19 Balanced accuracy taux de sous-estimation et ROC AUC du modèle de prédiction des défaillances après différents ajustement du seuil

4.2.1.2 Mesure des résultats « Métier » en termes de performance : Choix de la variable cible

Dans le cas d'une cible « Statut juridique », les résultats sont les suivants :

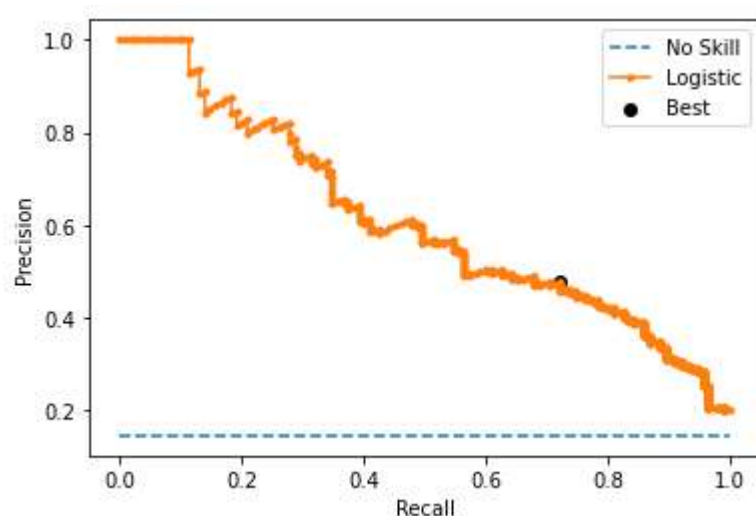


Figure 20 Courbe Précision Rappel du modèle de prédiction des défaillances "Statut juridique"

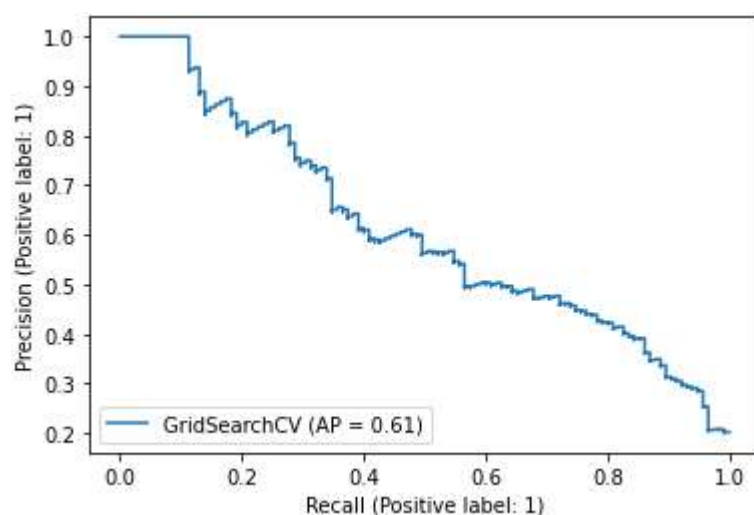


Figure 21 Courbe Précision Rappel et AP du modèle de prédiction "Statut juridique"

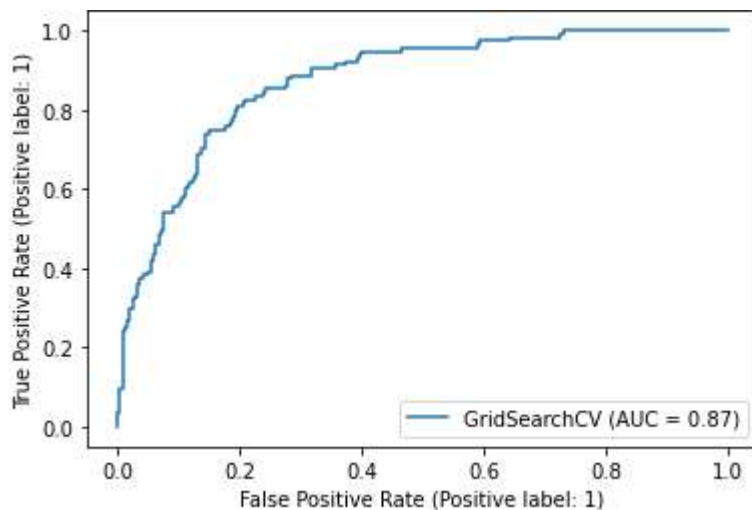


Figure 22 Courbe ROC et AUC du modèle de prédiction "Statut juridique"

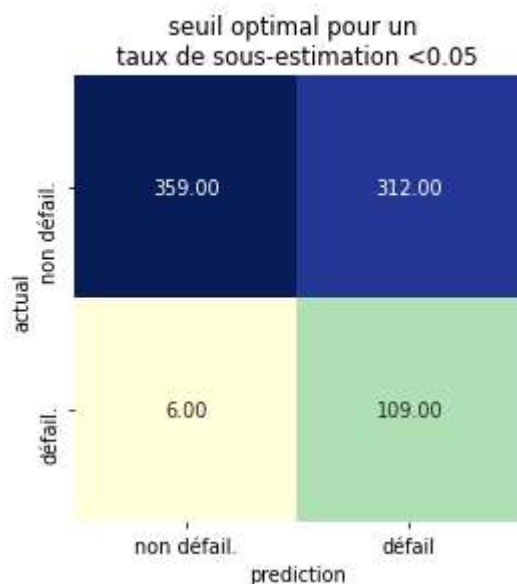


Figure 23 Nombre maximum de défaillances prédites en 2022 avec le statut juridique

model_name	taux de sous-estimation	ROC AUC
Signaux Faibles	0.593000	0.580000
Xgboost seuil par défaut	0.217391	0.873609
Xgboost seuil / f score	0.269565	0.873609
Xgboost seuil / taux de sous-estimation	0.052174	0.873609

Figure 24 supériorité des mesures des performances des prédictions de défaillances XGBoost seuil / taux de sous-estimation

Dans le cas d'une cible « Procédure Collective », les résultats sont les suivants :

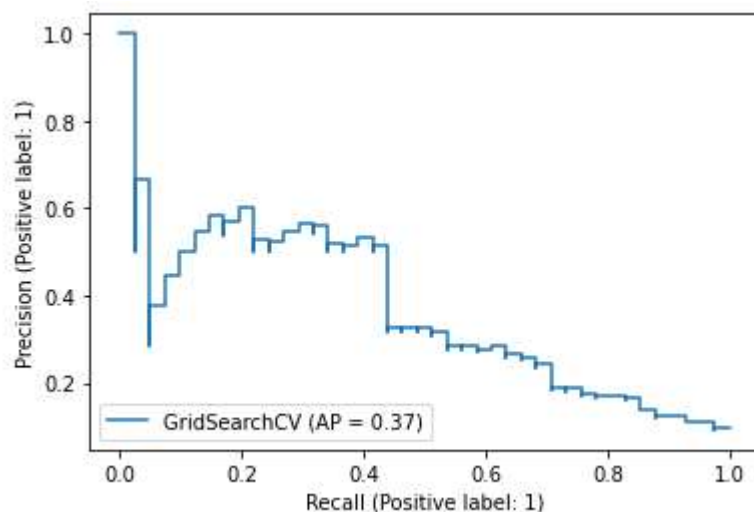


Figure 25 Courbe de Précision Rappel et AP du modèle de prédiction "Procédure Collective"

4.2.2 Mesure des résultats en termes de volume des prédictions des défaillances

Le tableau suivant croise les prédictions de défaillances d'entreprises avec les procédures collectives connues.

Dans le cas du seuil de classification défini par défaut, les résultats sont les suivants :

Défaillances prédites	Nombre d'entreprises
Non défailtantes	569
Proc. Col. Non	558
Proc. Col. Oui	11
Défaillantes	217
Proc. Col. Non	181
Proc. Col. Oui	36
(vide)	
(vide)	
Total général	786

Dans le cas du seuil de classification calculé pour minimiser le taux de sous-estimation, les résultats sont les suivants :

Défaillances prédites	Nombre d'entreprises
Non défailante	365
Proc. Col. Non	364
Proc. Col. Oui	1
Défaillante	421
Proc. Col. Non	375
Proc. Col. Oui	46
(vide)	
(vide)	
Total général	786

4.3 Discussion

Ce chapitre comporte la discussion d'une part des aspects « Machine Learning » et d'autre part des aspects « Métier » de la recherche menée dans l'approche empirique.

4.3.1 Discussion des aspects « Machine Learning » : Utilisation de la mesure du taux de sous-estimation d'un modèle et méthode d'ajustement d'un seuil de classification en fonction du taux de sous-estimation

Le lecteur vérifie deux hypothèses « Machine Learning » dans ce chapitre : l'utilisation de la mesure du taux de sous-estimation d'un modèle et la méthode d'ajustement d'un seuil de classification en fonction du taux de sous-estimation.

La discussion de ces deux hypothèses « Machine Learning » nécessite de rappeler les bases des mesures en Statistiques et leurs différences d'application en Médecine et en Machine Learning.

Puis la discussion aborde les conclusions sommaires, se poursuit avec la pertinence, la qualité, les implications de la recherche et se termine en évoquant les évolutions possibles de la recherche.

4.3.1.1 Différences des mesures statistiques en Médecine et en Machine Learning

Les mesures statistiques sont traditionnellement utilisées en Médecine. Elles diffèrent en Machine Learning, bien que leur objectif soit commun : évaluer une prédiction ou un test.

4.3.1.1.1 Mesures statistiques en Médecine et en Machine Learning

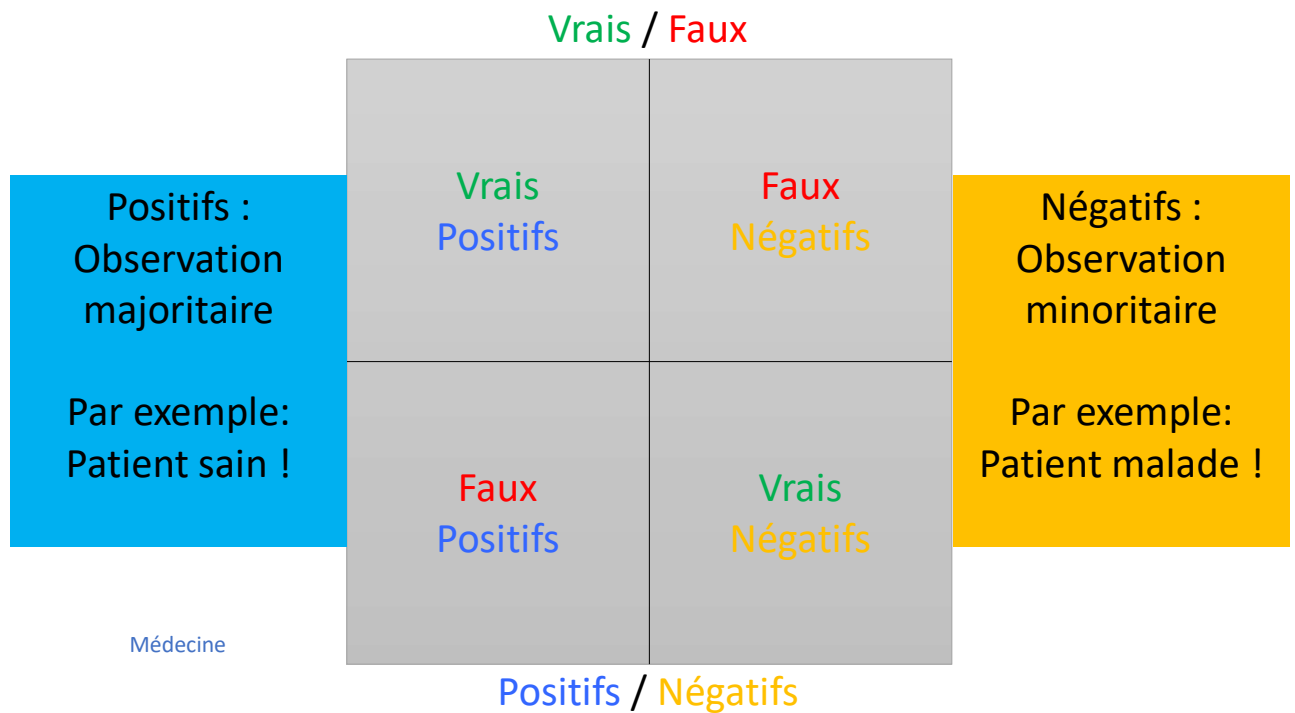


Figure 26 Matrice de confusion en Médecine

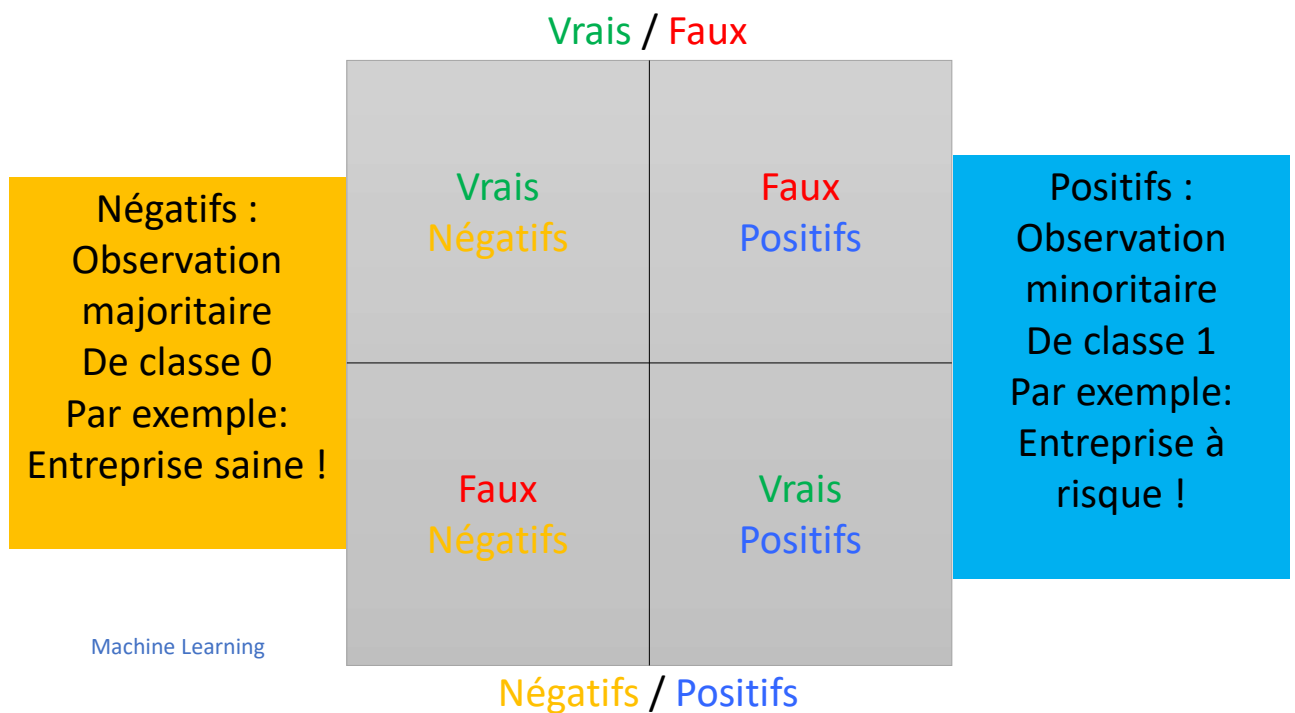


Figure 27 Matrice de confusion en Machine Learning

Les deux matrices sont inversées !

Les observations positives et négatives sont opposées.

4.3.1.1.2 Objectifs des prédictions en Médecine et en Machine Learning

Le but de la recherche en Médecine est d'éviter l'erreur de type I, à savoir, l'erreur provenant du rejet d'une hypothèse qui est vraie en réalité.

Par exemple, éviter de ne pas détecter une pathologie.

Mathématiquement, cela se traduit par diminuer le False Positif Rate $FPR = \frac{FP}{N}$

Un raisonnement similaire a été mené pour transposer et appliquer ces règles au Machine Learning dans la présente approche empirique.

Cela conduit, aux règles suivantes en prédiction des défaillances des entreprises :

Eviter de ne pas détecter une entreprise à risque de défaillance.

Mathématiquement, cela implique de transposer la règle appliquée en Médecine, décrite ci-dessus et conduit à :

Diminuer le False Negative Rate $FNR = \frac{FN}{P}$

Ce qui revient à éviter l'erreur de type II, à savoir, l'erreur provenant de l'acceptation d'une hypothèse qui est fausse en réalité. Cette erreur est aussi appelée « type II error », « miss », « underestimation » en anglais. (Wikipedia, s.d.).

En conséquence, le Machine Learning doit chercher à diminuer l'« underestimation » ou taux de sous-estimation en français.

Cette nouvelle règle a été appliquée dans cette approche empirique de prédiction des défaillances des entreprises.

4.3.1.2 Conclusions sommaires

Les conclusions portent sur les prédictions réalisées et les performances de celles-ci. Elles permettent de valider les deux hypothèses « Machine Learning » portant sur le taux de sous-estimation.

4.3.1.2.1 Résultats de la prédiction avec les différents seuils de classification

L'approche empirique permet de comparer les résultats de différents seuils de classification. Trois seuils sont comparés : le seuil de classification défini par défaut à la valeur 0,5, le seuil calculé pour optimiser la mesure F-Score et le seuil calculé pour minimiser le taux de sous-estimation.

Les trois prédictions se basent sur le même modèle. Seul, le seuil de classification des probabilités change.

La comparaison des 3 matrices de confusion résultantes indique que seul le seuil optimal fixé pour minimiser le taux de sous-estimation permet de réduire le nombre de Faux Négatifs en

dessous de la barre des 10 entreprises. En effet, le nombre de Faux Négatifs dépasse la vingtaine pour les autres seuils.

La comparaison des performances des différentes classifications permet de constater une bonne tenue des performances dans le cas d'un ajustement du seuil à des fins de minimisation du taux de sous-estimation :

- Taux de sous-estimation inférieur à 5 %.
- Mesure de ROC AUC au-dessus de 0,80
- Taux de balanced accuracy au-dessus de 0,74.

En conclusion, seul le seuil calculé pour optimiser la mesure du taux de sous-estimation permet de diminuer drastiquement l'erreur de type II, à savoir les entreprises défaillantes non détectées en tant que telles (cf. figure 20)

De plus, ce seuil calculé pour minimiser la mesure du taux de sous-estimation a aussi l'avantage de générer une très bonne performance en terme de ROC AUC, ainsi qu'en terme de balanced accuracy.

4.3.1.2.2 Vérification des hypothèses des problématiques de recherche

Le seuil de classification calculé pour minimiser la mesure du taux de sous-estimation est un seuil efficace de prédiction des défaillances, car la matrice de confusion résultante offre de bonnes performances, non seulement, en taux de sous-estimation, mais aussi en rappel et balanced accuracy.

Le seuil proposé par défaut ainsi que le seuil calculé pour optimiser la mesure F-Score ne sont pas de bons seuils de classification de prédiction des défaillances.

L'hypothèse est vérifiée.

4.3.1.3 Pertinence de la recherche

Il n'y a pas de mesure des FN dans la littérature.

Or les représentants du « Métier » ont besoin de mesurer cette performance d'un modèle. Elle était mesurée en statistiques par l'erreur de type II.

Il y a donc un besoin de recherche dans ce domaine.

L'erreur de type II (ou le taux de sous-estimation), est un problème persistant des modèles de prédiction des défaillances intégrant un seuil par défaut, ou même un seuil optimisant la mesure F-Score.

L'application du taux de sous-estimation, conjuguée à la mesure du rappel, est un sujet de recherche pertinent.

Cette recherche propose d'appliquer cette méthode en Machine Learning.

4.3.1.4 Qualité de la recherche

La qualité de cette recherche est discutée ici sur les trois aspects : validité des résultats, pertinence et défaut de la méthode.

4.3.1.4.1 Validité des résultats

Le tableau croisé dynamique croisant les prédictions de défaillance avec la variable explicative la plus importante à savoir les Bénéfices et Pertes de la figure 17 permet d'établir la justesse de la classification des entreprises à risque de défaillance.

4.3.1.4.1.1 Justesse du nombre de Faux Négatifs

Les FN sont au nombre de 6 dans l'approche empirique. Leur rubrique de Bénéfices ou pertes totales est proche de 0 contrairement aux TN, au nombre de 359, dont la même rubrique est largement positive. Ceci prouve que l'algorithme a bien séparé les entreprises défaillantes (avec des pertes) des entreprises non défaillantes (sans pertes).

La mesure du taux de sous-estimation est compatible avec une prédiction à horizon supérieur à 12 mois, alors que d'autres mesures intégrant les FP (accuracy, précision), ne le sont pas.

En effet, les FN concernent des entreprises qui ont connu des défaillances dans le passé : donc le modèle devrait avoir toutes les données pour les prédire et les vérifier. Le modèle doit avoir de très bonnes performances sur les FN et plus particulièrement sur le taux de sous-estimation.

4.3.1.4.1.2 Justesse du nombre de Faux Positifs

Dans l'approche empirique effectuée, le nombre de FP est de 312 (cf. fig.17 précédente).

Dans le tableau précédent, une rubrique Bénéfices / Pertes totales négative de -1, 116 Ke apparaît pour les FP. Ce total est négatif comme pour les 109 défaillances avérées et au contraire des 365 autres entreprises prédites non défaillantes (pour lesquelles ce total est largement positif).

Cela prouve que ces FP sont à risque de défaillances dans le futur. Cette mesure ne doit pas être prise en compte comme une erreur de prédiction, mais comme une vraie prédiction en attente de confirmation sur les comptes suivants.

En effet, les FP recouvrent aussi des prédictions dans le futur à 18 mois d'horizon, c'est à dire des entreprises à risque de défaillances dans les 18 mois, mais non encore enregistrées comme telles dans le dataset. En effet le dataset ne comporte pas encore les données du futur (le statut juridique « Inactif » de l'entreprise à risque) pour vérifier la future défaillance des entreprises détectées ainsi. Ces entreprises sont classées alors FP en attendant les comptes de l'année suivante ou des 2 années suivantes durant lesquelles la défaillance risque fort de survenir.

4.3.1.4.1.3 Justesse du seuil de classification calculé pour minimiser le taux de sous-estimation

Le nombre de FP (6) est bien inférieur dans le cas du seuil minimisant le taux de sous-estimation que dans le cas du seuil par défaut (24) ainsi que dans le cas du seuil calculé pour optimiser le F-Score (33).

De plus, dans le cas du seuil minimisant le taux de sous-estimation, les autres performances du rappel et de la balanced accuracy sont maintenues élevées.

Les résultats sont donc valides.

4.3.1.4.2 Pertinence de la méthode

La méthode a consisté à comparer les résultats de trois seuils de classification de prédiction des défaillances sur le même dataset et le même modèle, en faisant varier uniquement le seuil. Il prend alors les 3 valeurs suivantes : le seuil par défaut, le seuil F-Score optimum et le seuil minimisant le taux de sous-estimation.

Les critères d'évaluation ont été le nombre de FP, le rappel et la mesure de la balanced accuracy.

Cette méthode de comparaison des résultats prouve l'efficacité du seuil minimisant le taux de sous-estimation par rapport aux deux autres seuils.

4.3.1.4.3 Défauts de la méthode

La méthode de comparaison utilisée ne présente pas de défaut majeur.

4.3.1.5 Implications pour le domaine de recherche

4.3.1.5.1 Comparaison avec les études précédentes

La littérature ne rapporte aucune étude à ce sujet. Cette approche empirique est novatrice.

4.3.1.5.2 Implications pour le domaine

L'utilisation du taux de sous-estimation implique un changement important dans les méthodes d'évaluation des modèles de prédiction des défaillances. Des modèles initialement mesurés performant, ne sont pas évalués comme tels avec ce taux de sous-estimation.

Par ailleurs, cette mesure pourrait faciliter les échanges avec les représentants du « Métier », ces derniers pouvant exprimer leurs exigences suivant une nouvelle mesure : le taux de sous-estimation.

L'ajustement du seuil de classification des défaillances calculé pour minimiser le taux de sous-estimation a pour conséquence d'ajuster le volume de la liste des entreprises détectées à risque de défaillances. Il implique que le « Métier » est libre de fixer le taux de sous-estimation souhaité et donc les performances du modèle attendues. Une autre implication est que le « Métier » est aussi libre de fixer le volume de la liste des entreprises détectées à risque de défaillance qu'il souhaite traiter.

4.3.1.5.3 Perspectives de recherche

4.3.2 Discussion des aspects « Métier »

Le lecteur vérifie ici une hypothèse « Métier » dans ce chapitre : la caractérisation de la défaillance d'entreprise par la cessation d'activité.

La discussion de cette hypothèse « Métier » commence par aborder les conclusions sommaires, se poursuit avec l'étude de la pertinence, la qualité, les implications de la recherche et se termine en évoquant les évolutions possibles de la recherche.

4.3.2.1 Conclusions sommaires

Les conclusions portent d'une part sur le dataset d'origine et d'autre part sur les prédictions réalisées. Elles permettent de valider la première hypothèse « Métier ».

4.3.2.1.1 Conclusions intermédiaires concernant les données d'origine

La notion de procédure collective est définie par l'état de cessation de paiement.

La cessation d'activité a pour origine de multiples causes. Les plus nombreuses sont afférentes à la liquidation.

Selon les données financières et comptables étudiées, le nombre de procédures collectives est très largement inférieur à celui-ci des cessations d'activité.

La notion de procédure collective ne recouvre pas les nouvelles procédures amiables dont le nombre est largement supérieur.

4.3.2.1.2 Conclusion finale concernant la prédiction des défaillances d'entreprise

4.3.2.1.2.1 Résultats de la prédiction avec la cible « statut juridique »

Le modèle de prédiction des défaillances retenu est basé sur la cible du statut juridique (Actif versus Inactif). La courbe de mesure Précision-Rappel suivante indique visuellement son efficacité bien supérieure à un algorithme de type « No Skill » (réalisant une prédiction aléatoire).

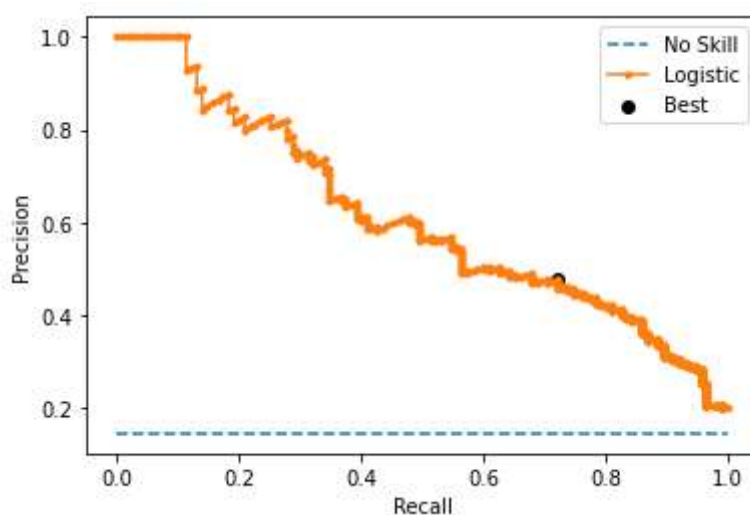


Figure 28 Courbe Precision Rappel du modèle de prédiction "Statut juridique"

La mesure de l'AP du modèle de prédiction des défaillances « Statut juridique » est satisfaisante : 0,61.

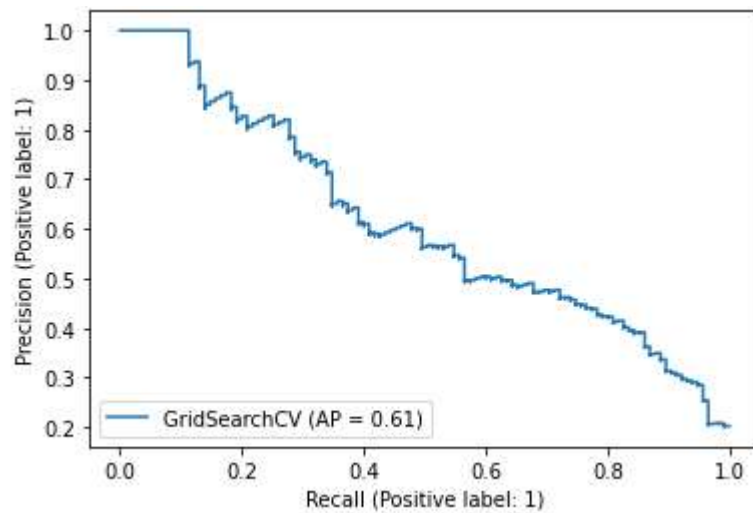


Figure 29 Courbe Precision Rappel du modèle de Prédiction "Statut juridique" et AP

La mesure de l'AUC ROC du modèle de prédiction des défaillances « Statut juridique » est encore meilleure : 0,87 (cf. courbe ROC suivante)

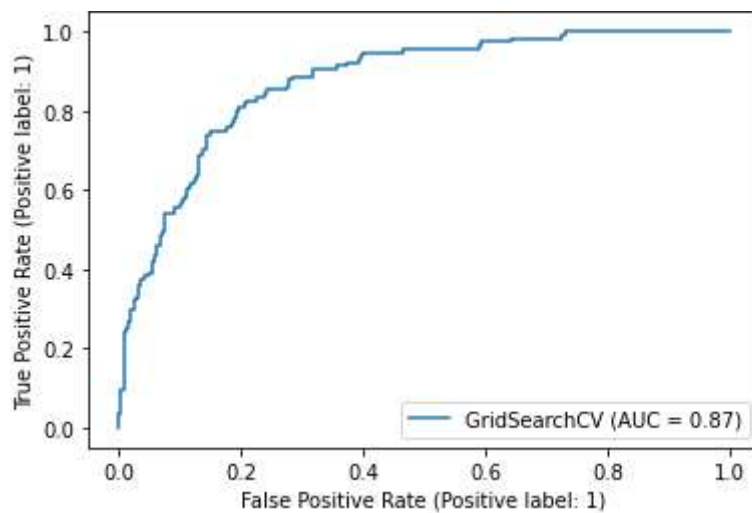


Figure 30 Courbe ROC et AUC du modèle de prédiction "Statut juridique"

Les résultats sur un échantillon de test de 786 entreprises, à la date du 21/08/2021 sont :

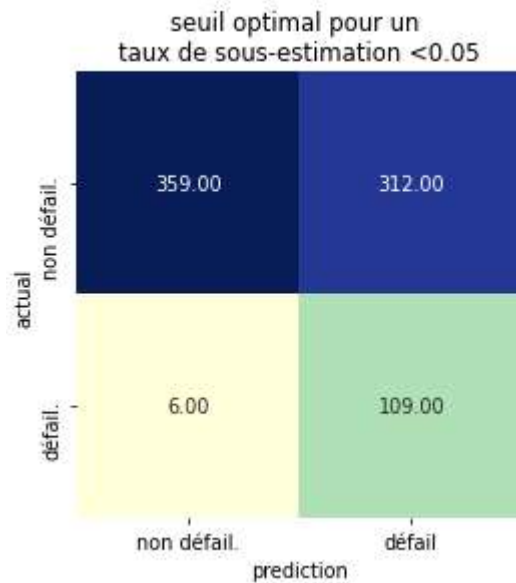


Figure 31 Nombre maximum de défaillances prédites en 2022 avec le statut juridique

Cette matrice de confusion se lit de la façon suivante :

- Les prédictions confirmées par la réalité sont les suivantes :
 - Nombre d'entreprises non-défaillantes confirmées : 359
 - Nombre d'entreprises défaillantes confirmées 109
- Les prédictions erronées sont les suivantes :
 - Nombre d'entreprises défaillantes non détectées 6
- Les « vraies prédictions » sont les suivantes :
 - Nombre d'entreprises non-défaillantes à ce jour mais détectées à fort risque de défaillance en 2022¹⁰ : 312

Performances du modèle de prédiction de défaillance XGBoost « Statut juridique »:

- Un seuil de classification des défaillances exigeant (0.29) a été utilisé afin d'obtenir un taux de sous-estimation inférieur à 5 %.
- Le modèle obtient un taux de sous-estimation de 0,0522 et une mesure de ROC AUC de 0,874.
- Les performances du modèle de prédiction de défaillance XGBoost « Statut juridique » sont bien supérieures aux performances de Signaux Faibles.

model_name	taux de sous-estimation	ROC AUC
Signaux Faibles	0.593000	0.580000
Xgboost seuil par défaut	0.217391	0.873609
Xgboost seuil / f score	0.269565	0.873609
Xgboost seuil / taux de sous-estimation	0.052174	0.873609

¹⁰ Prévision du risque pour la période du 01/01/2021 au 30/06/2022, durant laquelle 75% des défaillances prévues sont attendues

Figure 32 supériorité des mesures des performances des prédictions de défaillances XGBoost seuil / taux de sous-estimation

4.3.2.1.2.2 Résultats de la prédiction avec la cible « procédure collective »

Le modèle de prédiction basé sur la cible « procédure collective » n'est pas efficace :

La mesure de la précision ne dépasse guère 0,2 et les performances d'un modèle de type « No Skill » (réalisant des prédictions aléatoires).

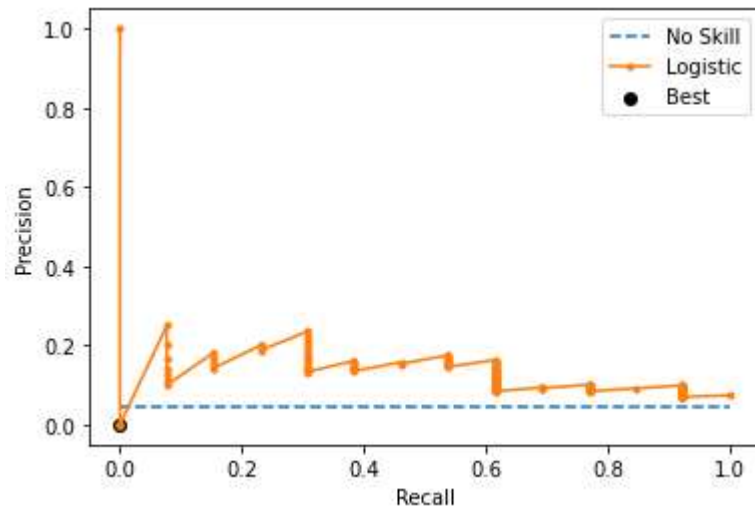


Figure 33 Courbe 1 Précision Rappel du modèle de prédiction des défaillances "procédure collective" abandonné

La mesure de l'AP du modèle de prédiction des défaillances « procédure collective » est médiocre : 0,15.

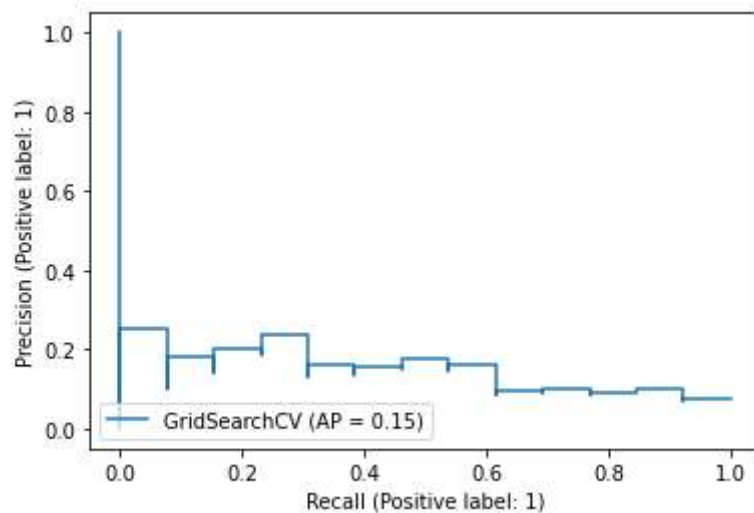


Figure 34 Courbe 2 Précision Rappel du modèle de prédiction des défaillances "procédure collective" abandonné

La mesure de l'AUC ROC (0,80) du modèle de prédiction des défaillances « procédure collective » est moins bonne que celle du modèle de prédiction des défaillances « Statut juridique » (0,87).

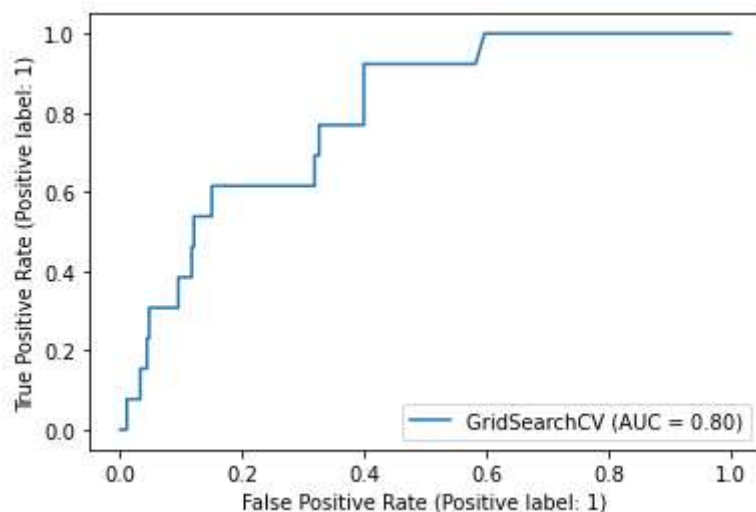


Figure 35 Courbe ROC et AUC du modèle de prédiction 'Procédure collective'

Le modèle de prédiction « Procédure collective » n'ayant pas obtenu de performance satisfaisante, la matrice de confusion n'est pas significative.

4.3.2.1.2.3 Vérification des hypothèses des problématiques de recherche

La cessation d'activité est une cible efficace de prédiction des défaillances, car le modèle révèle de bien meilleures performances en termes de rappel, de taux de sous-estimation, d'AP et de ROC AUC.

L'entrée en procédure collective n'est donc pas une cible efficace de prédiction des défaillances.

L'hypothèse de caractérisation de la défaillance par le statut juridique « Inactif » est vérifiée.

4.3.2.2 Pertinence de la recherche

L'état des lieux en prédiction datant de 2004 suggérait différentes caractérisations de la défaillance, sans avoir été étudiée par la suite (Refait-Alexandre, 2004).

Or, de plus en plus de défaillances d'entreprises ont lieu en dehors de la connaissance d'une procédure collective. Les prévisions de défaillances décrites dans la littérature perdent, ainsi, de leur pertinence.

Ce nouveau modèle de prédiction répond bien à la nouvelle problématique des cessations d'activité hors procédures collectives en générale et à la montée en puissance des procédures amiables en particulier.

Il donne une bonne indication des risques de défaillances des entreprises.

4.3.2.3 Qualité de la recherche

La qualité de cette recherche est discutée ici sur les trois aspects : validité des résultats, pertinence et éventuel défaut de la méthode.

4.3.2.3.1 Validité des résultats

Les mesures utilisées pour valider les résultats sont des mesures reconnues.

La mesure de l'AUC ROC est supérieure dans le cas du modèle de prédiction des défaillances ayant pour cible le « statut juridique » à celle du modèle de prédiction des défaillances ayant pour cible la « procédure collective ».

La courbe Précision Rappel de chacune, fait apparaître une mesure de Précision satisfaisante dans le cas du modèle de prédiction des défaillances ayant pour cible le « statut juridique » et une mesure de Précision médiocre dans le cas du modèle de prédiction des défaillances ayant pour cible la « procédure collective ».

Les résultats de cette recherche sont donc valides.

4.3.2.3.2 Pertinence de la méthode

La méthode a consisté à comparer les résultats des deux modèles de prédiction des défaillances sur le même dataset, en faisant varier uniquement la variable cible, à savoir le « Statut juridique » versus la « procédure collective ».

Cette méthode de comparaison des résultats prouve l'efficacité de la cible « Statut juridique » par rapport à la cible « Procédure collective ».

4.3.2.3.3 Eventuel défaut de la méthode

La méthode de comparaison utilisée ne présente pas de défaut.

4.3.2.4 Implications pour le domaine de recherche

4.3.2.4.1 Comparaison avec les études précédentes

Aucune publication précédente ne fait état de la cible « statut juridique » en prédiction des défaillances des entreprises françaises. Seule, une publication concernant la prédiction des Spin-offs belges en faillite ayant cessé leur activité est mentionnée dans la revue de littérature. Mais elle ne précise pas exactement sa cible.

4.3.2.4.2 Implications pour le domaine

Le changement de cible de prédiction des défaillances au profit du statut juridique implique deux apports pour le domaine de recherche. D'une part, l'amélioration de la qualité de telles prédictions offre une pertinence des résultats accrue. D'autre part, le caractère opérationnel de telles prédictions, malgré l'absence des informations concernant les procédures amiables, maintient la pertinence de ces prédictions à l'avenir.

4.3.2.4.3 Perspectives de recherche

4.4 Gestion de projet

La gestion de projet a été allégée du fait de la taille réduite du POC.

4.4.1 Finalités du Proof Of Concept

Pour mémoire, ce Proof Of Concept (POC), a pour objectif de valider la méthodologie de la prédiction des défaillances des entreprises, avant passage à l'échelle.

4.4.2 Liste des tâches à effectuer

Cinq phases de travaux ont été nécessaires : l'écoute du besoin, le cadrage de la solution, la conception du dataset, la conception et l'entraînement du modèle et enfin la réception du POC.

4.4.3 Ordonnancement des tâches

L'ordonnancement des tâches du POC a été le suivant :

- Ecoute du besoin auprès du service client le 26/03/2021
- Cadrage de la solution du 26/03 au 07/05 : 18 j
- Conception du dataset du 07/05 au 28/06 : 15 j
- Entraînement du modèle Xgboost du 16/06 -05/08 : 20 j
- Transmission du POC à DGA / MI le 15/09/2021

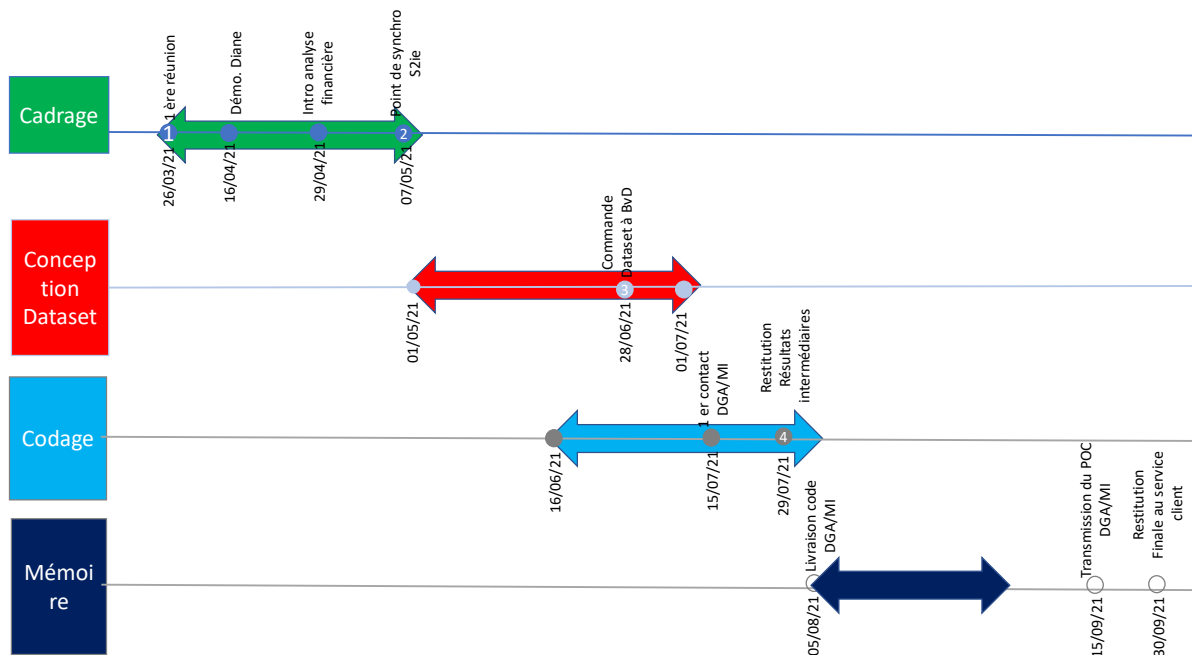
NB : une seule personne a constitué les ressources du POC, à temps partiel (2j / semaine) jusqu'au 01/06/21, puis à temps plein.

4.4.4 Durée du projet

Du 26/03/2021 au 30/09/2021

4.4.5 Jalons & livrables

Les jalons et phases du POC figurent sur le graphique suivant :



Le présent mémoire et les différentes versions du code réalisé constituent les livrables du POC.

4.4.6 Outils de suivi et de contrôle

Différentes réunions de suivi se sont déroulées tout au long du POC :

- Une réunion hebdomadaire avec le chef du projet POC
- Une réunion bimensuelle avec le service client

4.5 Réalisation d'un budget

Le principal coût du POC a porté sur l'achat du Dataset auprès du fournisseur de données : Diane / Bureau van Dijk.

5 Conclusion

L'erreur du type II, comme le nombre d'observations Faux Négatifs, donne une bonne mesure des performances d'un modèle de prédiction des défaillances. Ce sujet est particulièrement étudié en Médecine pour vérifier l'exactitude des diagnostics.

La recherche menée pour cette approche empirique a conduit à concevoir une nouvelle mesure des performances appelée taux de sous-estimation. Elle caractérise l'erreur de type II. Un bon modèle voit son taux de sous-estimation minimisé en dessous de 5% par exemple.

De plus, cette mesure peut être habilement utilisée pour définir un seuil de classification adéquat. Cette approche empirique fournit un exemple d'ajustement du seuil de classification en fonction d'un taux de sous-estimation à ne pas dépasser.

Cet ajustement permet, d'une part, de respecter des exigences de qualité et, d'autre part, d'ajuster le volume des résultats en fonction des attentes de la prédiction.

Bibliographie

- République Française Data.gouv.fr. (2021, août 07). *Signaux Faibles - Outil de détection des entreprises en risque de défaillance*. Consulté le août 12, 2021, sur Data.gouv.fr: <https://www.data.gouv.fr/fr/organizations/signaux-faibles/>
- Alfocea, A. (2021, avril 7). Comment le Machine Learning permet de détecter la fraude bancaire. *Management & Data Science*, 5(4). doi:<https://doi.org/10.36863/mds.a.16671>
- altares. (2020, décembre 17). *Comment prévoir les défaillances d'entreprises et prévenir les impayés ?* Consulté le août 18, 2021, sur DATA altares: <https://www.altares.com/fr/blog/2020/12/17/comment-prevoir-les-defaillances-dentreprises-et-prevenir-les-impayes/>
- Altman, E. (1968, septembre). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. (Wiley, Éd.) *The Journal of Finance*, 23(4), pp. 589-609. doi:<https://doi.org/10.2307/2978933>
- Altman, E., Marco, G., & Varetto, F. (1994, mai). Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). *Journal of Banking & Finance*, 18(3), 505-529. doi:[https://doi.org/10.1016/0378-4266\(94\)90007-8](https://doi.org/10.1016/0378-4266(94)90007-8)
- aquila data enabler. (2020, octobre 29). *INTERPRÉTABILITÉ DES MODÈLES DE MACHINE LEARNING*. Consulté le août 16, 2021, sur aquila data enabler: <https://www.aquiladata.fr/insights/interpretabilite-des-modeles-de-machine-learning/>
- Balboni, A., Boulegue, A., Mirlicourtois, A., Passet, O., & Paturel, P. (2021, juillet). Conjoncture et prévisions pour l'entreprise. (X. Previsis, Éd.) Récupéré sur <https://www.xerfi.com/flash/Xerfi-Previsis.pdf>
- Banque de France. (1995). Détection précoce des défaillances d'entreprises à partir des documents comptables. *Bulletin de la Banque de France supplément "Etudes"*. Consulté le août 20, 2021, sur http://www.assurances-chomage.fr/medias/articles/docs/detection_defaillances_entreprises_partir_documents_comptables-66-2.PDF
- Banque De France Eurosystem. (2020, octobre 08). *Rapport de l'Observatoire des délais de paiement*. Consulté le août 18, 2021, sur Banque De France Eurosystem: <https://publications.banque-france.fr/liste-chronologique/rapport-de-lobservatoire-des-delaix-de-paiement>

- Banque De France Eurosystem. (2021, août 11). *Les défaillances d'entreprise*. Consulté le Août 12, 2021, sur Banque De France Eurosystem Webstat: http://webstat.banque-france.fr/fr/browseSelection.do?node=5384339&SERIES_KEY=DIREN.M.FR.DE.UL.DF.07.N.ZZ.TT&SERIES_KEY=DIREN.M.FR.DE.UL.DF.03.N.BE.PM
- Banque de France Eurosystem. (2021, août 11). *STAT INFO*. Consulté le août 12, 2021, sur <https://www.banque-france.fr/statistiques/defaillances-dentreprises-jul-2021>
- Barboza, F., Kimura, H., & Altman, E. (2017, october). Machine learning models and bankruptcy prediction. *Expert Systems with applications*, 83, 405-417. doi:<https://doi.org/10.1016/j.eswa.2017.04.006>
- Barboza, F., Kimura, H., & Altman, E. (2017). *Machine Learning Models and Bankruptcy Prediction* (Vol. 83). Elsevier Expert Systems with Applications. doi:10.1016/j.eswa.2017.04.006
- Bardos, M., & Zhu, W. (1997). Comparaison de l'analyse discriminante linéaire et des réseaux de neurones. Application à la détection de défaillance d'entreprises. (S. F. statistique, Éd.) *Revue de Statistique Appliquée*, 45(4), 65-92. Récupéré sur http://www.numdam.org/item/RSA_1997__45_4_65_0/
- Beaver, W. (1966). Financial Ratios As Predictors of Failure. (Wiley, Éd.) *Journal of Accounting Research*, 4, *Empirical research in Accounting : Selected Studies 1966*, pp 71 - 111. doi:<https://doi.org/10.2307/2490171>
- Ben Jabeur, S., & Fahmi, Y. (2013). Prédiction de la défaillance des entreprises : une approche de classification par les méthodes de Data-Mining. *CAIRN.INFO Matières à réflexion*, 30(2013/4), 31 à 45. doi:<https://doi.org/10.3917/g2000.304.0031>
- Branco, P., Torgo, L., & Ribeiro, R. (2015). *A Survey of Predictive Modelling under Imbalanced Distributions*. Cornell University. Consulté le août 24, 2021, sur <https://arxiv.org/abs/1505.01658>
- Brownlee, J. (2020). *Imbalanced Classification with Python Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*. Machine Learning Mastery. Consulté le août 24, 2021, sur <https://machinelearningmastery.com/imbalanced-classification-with-python/>
- Bureau van Dijk. (s.d.). diane. Consulté le avril 2021, sur https://www.bvdinfo.com/fr-fr/nos-produits/donnees/national/diane?gclid=Cj0KCQjwvaeJBhCvARIsABgTDM5Y0lcuqCLjla48T9-iugEjR_ReMVkRN9itrgAYpX3PAxvpgfYT4HcaAo99EALw_wcB
- Camilleri, P., Joli, A., Viers, V., Coufour, F., & Ninucci, C. (2021, may 25). *Signaux-Faibles / Documentation*. Consulté le août 16, 2021, sur GitHub: <https://github.com/signaux-faibles/documentation/blob/master/description-donnees.md>
- CCI Paris Ile De France. (2019). *Publicité des comptes annuels de certaines sociétés*. Consulté le août 17, 2021, sur CCI Paris Ile De France Entreprises: <https://www.entreprises.cci->

paris-idf.fr/web/reglementation/nos-produits/docpratic/actualites-juridiques/publicite-comptes-annuels


- Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal Of Artificial Intelligence Research*, 16, 321-357. doi:<https://doi.org/10.1613/jair.953>
- Chen, I. (2019, janvier 2). *Basic Ensemble Learning (Random Forest, AdaBoost, Gradient Boosting)- Step by Step Explained*. Consulté le août 16, 2021, sur Towards data science: <https://towardsdatascience.com/basic-ensemble-learning-random-forest-adaboost-gradient-boosting-step-by-step-explained-95d49d1e2725>
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. (a. A. Machinery, Éd.) *ACM Journals*, 785-794. doi:<https://dl.acm.org/doi/abs/10.1145/2939672.2939785>
- data.gouv.fr, R. F. (s.d.). Base Sirene des entreprises et de leurs établissements (SIREN, SIRET). Consulté le juin 2021, sur <https://www.data.gouv.fr/fr/datasets/base-sirene-des-entreprises-et-de-leurs-etablissements-siren-siret/>
- Delcaillau, D. (2019). *Contrôle et transparence des modèles complexes en actuariat*. Mémoire d'actuariat, EURIA IMT Atlantique. Consulté le août 2021, sur <https://www.institutdesactuaires.com/se-documenter/memoire-d-actuariat-38?id=edbc96af9e65445f1c347bc48eaa1ba2>
- Deng, X., Li, M., Deng, S., & Wang, L. (2021, may 30). Hybrid gene selection approach using XGBoost and multi-objective genetic algorithm for cancer classification. China. Consulté le août 23, 2021, sur <https://arxiv.org/abs/2106.05841>
- Dietsch, M., & Gonzalez, O. (2020, février 21). *Les retards de paiement des clients impactent-ils la probabilité de défaillance des entreprises ?* Consulté le août 18, 2021, sur Banque De France Eurosystem: <https://publications.banque-france.fr/les-retards-de-paiement-des-clients-impactent-ils-la-probabilite-de-defaillance-des-entreprises>
- Durand, P., & Le Quang, G. (2021). What do bankruptcy prediction models tell us about. *CNRS Université Paris Nanterre Economix Working Paper*, 2021-2. Récupéré sur https://economix.fr/pdf/dt/2021/WP_EcoX_2021-2.pdf
- El Alaoui, I. (2014, janvier 27). *Les méthodes ensemblistes pour algorithmes de machine learning*. Consulté le août 16, 2021, sur OCTO Talks ! le blog des Octos: <https://blog.octo.com/les-methodes-ensemblistes-pour-algorithmes-de-machine-learning/>
- Elbaz, S. (2020, septembre 24). *Random Forest : Forêt d'arbre de décision- Définition et fonctionnement*. Consulté le août 16, 2021, sur DataScientest: <https://datascientest.com/random-forest-definition>
- ellisphere. (s.d.). *search & monitor by ellisphere.fr*. Récupéré sur ellisphere: <https://www.ellisphere.fr/>

- Ferire, L. (2017). *La validation de modèles financiers prédictifs de défaillance au travers d'une analyse des spin-offs belges en faillite*. mémoire de master en sciences de gestion, HEC-Ecole de gestion de l'Université de Liège, Surlemont, Bernard. Consulté le août 18, 2021, sur <https://matheo.uliege.be/handle/2268.2/3620>
- Fernandez, A., Garcia, S., Galar, M., Prati, R., Krawczyk, B., & Herrera, F. (2018). *Learning from Imbalanced Data Sets*. Springer Link. doi:<https://doi.org/10.1007/978-3-319-98074-4>
- Foglia, A., Marullo, P., & Marullo Reedtz, P. (1998). Multiple banking relationships and the fragility of corporate borrowers. (E. B. V., Éd.) *Journal of Banking and Finance*, 22(10-11), 1441-156. Consulté le août 25, 2021, sur <https://www.sciencedirect.com/science/article/abs/pii/S0378426698000582>
- France Stratégie Evaluer, Anticiper, Débattre, Proposer. (2020, décembre 14). *Les défaillances d'entreprises dans la crise Covid-19 : zombification ou mise en hibernation ?* Consulté le août 20, 2021, sur République Française Liberté Egalité Fraternité France Stratégie Evaluer, Anticiper, Débattre, Proposer: <https://www.strategie.gouv.fr/point-de-vue/defaillances-dentreprises-crise-covid-19-zombification-mise-hibernation>
- Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189-1232. doi: DOI: 10.1214/aos/1013203451
- George, S., Dehoux, A., & Liao, C. (2019, octobre 22). *SHAP : MIEUX COMPRENDRE L'INTERPRÉTATION DE MODÈLES*. Récupéré sur Aquila Data Enabler: <https://www.aquiladata.fr/insights/shap-mieux-comprendre-linterpretation-de-modeles/>
- GitHub / Signaux Faibles. (2021, march). *Signaux Faibles Forge des solutions Signaux Faibles*. Consulté le août 16, 2021, sur GitHub: <https://github.com/signaux-faibles>
- Guerin, D., & Branchu-Bord, H. (2021, août 19). *Guide pratique procédure de traitement des difficultés des entreprises*. Consulté le août 19, 2021, sur SELAS GUERIN et Associés Madataire judiciaire: <https://www.dominique-guerin.fr/info/missions/349/guide-pratique>
- He, H., & Ma, Y. (2013). *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley-IEEE Press. Consulté le août 24, 2021, sur <https://books.google.fr/books?hl=fr&lr=&id=CVHx-Gp9jzUC&oi=fnd&pg=PT9&dq=Imbalanced+Learning:+Foundations,+Algorithms,+and+Applications&ots=2iMqLoEudg&sig=qvFUjG9NdcawsNtzNus7IXNFawo#v=onepage&q=Imbalanced%20Learning%3A%20Foundations%2C%20Algorithms%2C%20>
- INPI. (2017, 03 29). *L'INPI lance l'ouverture des données issues du registre du commerce et des sociétés*. Consulté le août 20, 2021, sur INPI: <https://www.inpi.fr/fr/nationales/l-inpi-lance-l-ouverture-des-donnees-du-registre-du-commerce-et-des-societes>

- INSEE. (2019, novembre 13). *Définitions, méthodes et qualité Défaillance d'entreprise*. Consulté le août 13, 2020, sur Insee Institut national de la statistique et des études économiques Mesurer pour comprendre: <https://www.insee.fr/fr/metadonnees/definition/c1617>
- INSEE. (2021, juin 01). *Diffusion open data*. Consulté le juin 02, 2021, sur sirene.fr: <https://www.sirene.fr/sirene/public/static/open-data>
- Koehl, M. (2019). *La négociation en droit des entreprises en difficulté*. Université de Nanterre Parix X, droit. HAL archives-ouvertes.fr. Consulté le août 18, 2021, sur <https://tel.archives-ouvertes.fr/tel-02280411/document>
- Laurent, J. (2021). *Une chute d'activité inédite dans l'industrie manufacturière en 2020*. INSEE. Consulté le août 12, 2021, sur <https://www.insee.fr/fr/statistiques/5405962>
- Lecointre, G. (2011). *Le grand livre de l'Economie PME 2012* (éd. 2). (Gualino, Éd.)
- Légifrance le service public de la diffusion du droit de la République Française. (2014, octobre 18). *Décret n° 2014-1189 du 15 octobre 2014 relatif à l'allégement des obligations de publicité des comptes annuels des micro-entreprises*. Consulté le août 17, 2021, sur Légifrance: <https://www.legifrance.gouv.fr/loda/id/LEGIARTI000029597780/2014-10-18#LEGIARTI000029597780>
- Levratto, N. (2017). *Le processus de défaillance des entreprises*. article, Centre de recherche international pour le développement des PME Université Paris Nanterre. Consulté le août 20, 2021, sur <https://recherche-developpementpme.parisnanterre.fr/articles-et-auteurs/le-processus-de-defaillance-des-entreprises-590121.kjsp>
- Levratto, N., Carré, D., Zouikri, M., & Tessier, L. (2011). *La défaillance des entreprises Etude sur données françaises entre 2000 et 2010*. Laboratoire Economix, Observatoire des PME d'OSEO. La Documentation Française. Consulté le août 12, 2021, sur <https://www.parisnanterre.fr/publications/la-defaillance-des-entreprises-etude-sur-donnees-francaises-entre-2000-et-2010-392550.kjsp>
- Ly, A. (2019). *Algorithmes de machine learning en assurance : solvabilité, textmining, anonymisation et transparence*. Université Paris Est Marne-la-Vallée (UPEM), École doctorale 532 : Mathématiques et STIC (MSTIC). HAL archives-ouvertes.fr. Consulté le août 16, 2021, sur <https://tel.archives-ouvertes.fr/tel-02413664v2/document>
- Me Robine, A. (2021, juillet 20). *La procédure de sauvegarde : les étapes à connaître*. Consulté le août 18, 2021, sur Captain Contrat: <https://www.captaincontrat.com/fermeture/entreprise-en-difficulte/procedure-sauvegarde-me-beaubourg-avocats>
- Memon, N., Patel, S., & Patel, D. (2019). *Comparative Analysis of Artificial Neural Network and XGBoost Algorithm for PolSAR Image Classification*. Inde. doi:10.1007/978-3-030-34869-4_49

- Metzler, G. (2019). *Learning from imbalanced data : an application to bank*. LHC Laboratoire Hubert Curien (Saint Etienne). HAL INRIA. Consulté le août 15, 2021, sur <https://hal.inria.fr/tel-02318899v1>
- Ministère de l'économie des finances et de la relance Bercy Entreprises Infos. (2019, septembre 18). *Qu'est-ce que la cessation de paiement d'une entreprise ?* Consulté le août 18, 2021, sur [economie.gouv.fr: https://www.economie.gouv.fr/entreprises/cessation-paiement-entreprise](https://www.economie.gouv.fr/entreprises/cessation-paiement-entreprise)
- Ministère de l'économie, des finances et de la relance, Bercy Infos. (2019, octobre 18). *Entreprises en difficulté : tout savoir sur les procédures collectives*. Consulté le août 18, 2021, sur [economie.gouv.fr: https://www.economie.gouv.fr/entreprises/entreprises-difficulte-procedures-collectives](https://www.economie.gouv.fr/entreprises/entreprises-difficulte-procedures-collectives)
- Morde, V. (2019, April 8). XGBoost Algorithm: Long May She Reign! The new queen of Machine Learning algorithms taking over the world.... (T. Editors, Éd.) *Towards Data Science*. Consulté le août 15, 2021, sur <https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d>
- Ottou, P. (2017). *Méthodes d'apprentissage automatique appliquées au provisionnement ligne à ligne en assurance non-vie*. Mémoire de master , Université Paris IX Dauphine. Consulté le août 23, 2021, sur <https://www.institutdesactulaires.com/docs/mem/b22bc48b9f8858867c7653eb8921e3d8.pdf>
- Prépa Dalloz Lefebvre Dalloz. (2016, août 2016). *Procédures collectives : la confidentialité face à la presse*. Consulté le août 18, 2021, sur [Le petit juriste: https://www.lepetitjuriste.fr/procedures-collectives-confidentialite-face-a-presse/](https://www.lepetitjuriste.fr/procedures-collectives-confidentialite-face-a-presse/)
- Refait-Alexandre, C. (2004). *La prévision de la faillite fondée sur l'analyse financière de l'entreprise : un état des lieux* (Vol. 2004/1 no 162). La Documentation Française Economie & Prévision. Consulté le août 13, 2021, sur <https://www.cairn.info/revue-economie-et-prevision-2004-1-page-129.htm#:~:text=%C3%80%20un%20an%20de%20la,est%20alors%20%C3%A9gal%20%C3%A0%2077%25>.
- République Française. (2021, juillet). *Signaux Faibles Mieux cibler les interventions en remédiation de l'État vers les entreprises en difficulté*. Consulté le août 16, 2021, sur [beta.gouv.fr: https://beta.gouv.fr/startups/signaux-faibles.html](https://beta.gouv.fr/startups/signaux-faibles.html)
- Schwab, P.-N. (2019, mai 17). *Bercy analyse les données pour prévoir les défaillances d'entreprises*. Consulté le août 20, 2021, sur [Into the Minds: https://www.intotheminds.com/blog/bercy-donnees-defaillances-entreprises-faillites/](https://www.intotheminds.com/blog/bercy-donnees-defaillances-entreprises-faillites/)
- Steen, D. (2020, september 20). *Precision-Recall Curves Sometimes a curve is worth a thousand words - how to calculate and interpret precision-recall curves in Python*.

- Consulté le août 24, 2021, sur Doug Steen:
<https://medium.com/@douglassteen/precision-recall-curves-d32e5b290248>
- Suchier, H.-M. (2006). *Nouvelles contributions du boosting en apprentissage*. Université Jean Monnet Saint Etienne, Informatique. HAL archives-ouvertes.fr. Consulté le août 22, 2021, sur <https://tel.archives-ouvertes.fr/tel-00379539/document>
- T, A. (2020, octobre 19). *Algorithmes de Boosting – AdaBoost, Gradient Boosting, XGBoost*. Consulté le août 16, 2021, sur DataScientest: <https://datascientest.com/algorithmes-de-boosting-adaboost-gradient-boosting-xgboost>
- Tremblay, C. (2021, mai 30). *Imbalanced data et Machine Learning Comprendre et traiter les données déséquilibrées*. Consulté le août 15, 2021, sur Kobia Devançons l'avenir: <https://kobia.fr/imbalanced-data-et-machine-learning/>
- Varetto, F. (1998, octobre). Genetic algorithms applications in the analysis of insolvency risk. *Journal of Banking & Finance*, 22(10-11), 1421-1439. doi:[https://doi.org/10.1016/S0378-4266\(98\)00059-4](https://doi.org/10.1016/S0378-4266(98)00059-4)
- Wikipédia. (s.d.). *Crédit fournisseur*. Consulté le août 19, 2021, sur Wikipédia L'encyclopédie libre: https://fr.wikipedia.org/wiki/Cr%C3%A9dit_fournisseur
- Wikipédia L'encyclopédie libre. (2020, octobre 9). *Arbre de décision*. Consulté le août 16, 2021, sur Wikipédia L'encyclopédie libre: https://fr.wikipedia.org/wiki/Arbre_de_d%C3%A9cision
- Wikipédia L'encyclopédie libre. (2021, janvier 7). *Arbre de décision (apprentissage)*. Consulté le août 16, 2021, sur Wikipédia L'encyclopédie libre: [https://fr.wikipedia.org/wiki/Arbre_de_d%C3%A9cision_\(apprentissage\)](https://fr.wikipedia.org/wiki/Arbre_de_d%C3%A9cision_(apprentissage))
- Wikipédia. (s.d.). *Sensibilité et spécificité*. Récupéré sur Wikipédia, L'encyclopédie libre: https://fr.wikipedia.org/wiki/Sensibilit%C3%A9_et_sp%C3%A9cificit%C3%A9
- Wikistat.fr. (2016, janvier 21). *Agrégation de modèles*. Récupéré sur wikistat.fr: <https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-agreg.pdf>



IA School
18 rue du Dôme
92100 Boulogne-Billancourt
France

IA SCHOOL
L'ÉCOLE DE L'INTELLIGENCE ARTIFICIELLE