Les enjeux des décisions bancaires automatisées par l'IA : Réflexions méthodologiques et mise en place à grande échelle sur les dépassements débiteurs des comptes courants dans une banque

Pascal Damel – Hoai Le Thi An [1] – Van Tuan Pham

L.C.O.M.S. – Université de Lorraine

[1] Professeur Institut Universitaire de France I.U.F.

LA CHAIRE RIA&MIT

En formation

vous présente sa 3ème édition du Workshop

- Intelligence Artificielle et impacts organisationnels -

Résumé

La mise en place de l'IA dans les banques est un enjeu technique et stratégique. La stratégie s'inscrit dans des gains de productivité nécessaires pour maintenir les niveaux de rentabilité. La mise en place de l'IAE pose des difficultés techniques et méthodologiques. Cet article présente une expérience à grande échelle d'une mise en place de l'IA dans une banque pour automatiser avec des « machines learning » toutes les décisions de dépassements débiteurs sur les comptes courants des clients entreprises et personnes physiques. Nous avons dans une première phase été confrontés à la complexité des systèmes d'informations bancaires. Nous présentons ensuite les principaux choix méthodologiques possibles. Nous justifions dans cet article les arbitrages méthodologiques pris en suivant une approche originale hybride qui combine la méthode C.A.R.T. et S.V.M.. Nous avons d'autre part proposé avec succès des avancées méthodologiques sur les « machines Learning » en créant une méthode plus adaptée « C.S.B. -D.C.A. » pour des données asymétriques entre les acceptations de découverts et les refus. Les bases de données concernées dans cette étude comprennent effectivement plus d'acceptation que de refus.

1. Introduction

Les banques utilisent de plus en plus l'IA pour automatiser certaines décisions bancaires en vue de réduire les couts d'exploitation. Les dépassements débiteurs des comptes sont à la fois une source de profit avec des prises de commissions et une source de risque en absence de garantie sur les sommes prêtées en débit. La volumétrie importante des traitements

manuels des dépassements de compte génère des coûts d'exploitation importants. En effet, chaque dépassement débiteur sur un compte courant doit être validé par le chargé du compte. Beaucoup d'autorisations sont évidentes et récurrentes, d'autres nécessitent une réflexion approfondie et complexe qui dépend d'éléments multifactoriels (relation client – encours clients – type de client personne morale ou personne physique– le couple rentabilité/risque …).

Les banques entament largement cette révolution technologique et méthodologique avec les réflexions suivantes.

Quelles méthodologies utilisées ? C.A.R.T. – *Machine learning* M.L.– *deep learning* ?

Quelles technologies ?

Doit-on automatiser toutes les décisions et sous quelles conditions ?

Quelles sont les conséquences en matière de « supply chain » financière pour les entreprises et les nombreux impacts organisationnels pour l'entreprise.

Nous présenterons dans cet article les différentes réflexions et avancées que nous avons formalisées lors de la mise en place d'une I.A. en décisions de dépassement dans une banque européenne pour l'ensemble de la clientèle entreprises et personnes physiques avec près de 450 000 décisions.

2. Aspects épistémologiques et base de données

Ce type d'études s'inscrivent parfaitement dans des démarches positivistes. Les données sont des instruments d'induction logiques pour comprendre les décisions prises.

1.1. Les enjeux organisationnels des découverts bancaires

Les découverts bancaires sur les comptes courants sont des sujets d'actualité pour les consommateurs, les banques et les instances réglementaires européennes. Les comptes courants sont des produits bancaires sur lesquels les banques réalisent une grande partie de leurs marges commerciales. Les marges commerciales calculées avec des méthodes des taux de marché [5] [6] sur les comptes courants sont importantes à cause de leur taux créditeur proche de zéro et des commissions d'intervention très largement facturées lorsque le découvert autorisé est dépassé.

La banque dispose également d'un pouvoir discrétionnaire lorsque le client dépasse le seuil débiteur maximum. Elle peut refuser le découvert ou « *overdraft* » et par conséquent déclarer le client en cessation de paiement. Cette décision est évidemment humaine avec des niveaux hiérarchiques qui dépend du type de client et des montants concernés.

L'enjeu d'automatiser les découverts est une décision difficile et stratégique. Une mauvaise automatisation peut entrainer des risques légaux et opérationnels pour la banque et ses clients. La réussite de cette automatisation permet à la banque de gagner de nombreux

équivalents en temps plein. En cela, cette mise en place concerne également l'aspect ressources humaines. Selon le discours stratégique de la banque, les équivalents temps plein gagnés par l'I.A., permettent de les réaffecter vers des taches commerciales supplémentaires normalement créatrices de P.N.B. (produit net bancaire).

Les découverts bancaires sont complexes à appréhender puisque les clients sont différents. La banque doit également prendre au sérieux ses décisions, car le compte courant n'offre en général pas de garantie pour la banque en cas d'impayés. Ces impayés se traduisent comptablement par des provisions pour dépréciation.

Compte tenu des risques, les banques en général limitent les décisions d'automatisation sur des montants modestes (quelques centaines d'euros voire quelques milliers en fonction des types de clients).

1.2. Les bases de données

Pour commencer cette étude, le premier problème est de comprendre la complexité du système d'information de la banque et des contraintes réglementaires propres à l'environnement bancaire. Comme contrainte réglementaire nous avons respecté la règlementation I.F.R.S. [1] 8 sur la segmentation de la clientèle. Cette segmentation est bien évidemment juridique mais aussi une classification « business » des clients. Nous avons également utilisé la réglementation bâloise (accord de Bâle 4) qui régule le niveau des fonds propres des banques en fonction des risques (crédit -marché – liquidité – opérationnels) pris par la banque. Cette réglementation offre des systèmes de calcul de risque déjà fiabilisés. Le contrôle de gestion est aussi une source d'information sur la rentabilité /risque des clients. Le C.R.M. offre auusi aussi des données. Bien évidemment, il reste à appréhender la mécanique informatique qui réalise la traçabilité des dépassements des découverts.

Ces bases sont complexes à obtenir car l'environnement de ces données ne sont pas toutes structurées au sein d'un système d'information centralisé. Les banques doivent composer avec des systèmes informatiques différents dont certains datent des années 1970 (IBM400). Les données sont aussi cloisonnées dans différentes chaines informatiques pour répondre aux besoins spécifiques réglementaires et internes (Fonction comptable – fonction controlling – fonctions back office – marketing…). Apres un audit conséquent de la qualité des informations disponibles, nous pouvons caractériser les informations clients comme suit:

– Informations K.Y.C (Know you customer) sur les critères économiques, juridiques, fiscaux.
– Informations sur les risques financiers (Probabilité de défaut bâlois – informations comptables).
– Informations sur le contrôle de gestion ou le *Controlling* (Historique de rentabilité client et segment marketing).
– Informations sur le fonctionnement des comptes courants et joints (Epargne – nombre de crédits).
– Informations techniques liées à la décision d'accepter ou de refuser le découvert sur le compte courant.

[1] I.F.R.S. International Financial Report Standard

1.3. Quelles variables ?

Les milliers de variables disponibles et les quantités impressionnantes de données ne permettent pas d'automatiser les décisions à partir de techniques globales comme le « deep learning ».

Pour faciliter le travail, les données ont été présélectionnées. Nous avons retenu les variables d'expert issues de la concertation avec les managers des différents sources analytiques. Nous avons disposé de 432 966 décisions de dépassement dont 372 811 décisions favorables et 60 155 décisions défavorables réparties entre les différents segments de la clientèle. Ces variables ont fait l'objet de tests statistiques classiques comme les statistiques descriptives, les tests d'hypothèse (moyenne – variance – distribution statistique).

Ces tests ont pour objectif de mieux comprendre les décisions du processus d'approbation de crédit. Nous avons effectué trois analyses statistiques : test F (ANOVA), tests T (moyennes) et khi deux (test de saisonnalité) pour dissocier les variables entre les décisions. Alors que le test T permet d'expliquer l'importance des différences entre les moyennes des groupes, l'idée principale derrière l'ANOVA pour la sélection des caractéristiques est de tester la signification statistique de la contribution de chaque caractéristique à la caractéristique de la réponse. L'objectif de ces tests est d'identifier les variables pertinentes à prendre en compte dans le processus de prise de décision et d'exclure celles qui ne sont pas. Ces tests permettent de réduire les variables de plusieurs milliers à une volumétrie proche de la centaine. Les résultats de l'analyse des principales données sont présentés dans le tableau 1 sur un segment clientèle particulier en « retail premium».

Tableau 1 : tests statistiques

MAY2016 PR 2	MONTANT_AUTORISE_FORCAGE_modif	PROBABILITE_DEFAUT	GLOBAL_ASSETS	MT_REVENU_TOTAL	Nb_Doscred	Anciencpt	V11_UNI	V32_UNI	V51_UNI
TESTS STATS									
Average (moment 1)	303.9911213	470.7232222	10248.90734	102.4294222	2.549951503	5469.103275	24.7311412	2600.959213	59.91093903
STD (moment 2)	271.5266572	1231.564632	21343.04395	184.215403	1.564434571	3210.374885	41.41672473	8037.941906	42.66304453
SKEWNESS (moment 3)	0.879321625	6.011177878	7.494997096	3.849055523	1.696794808	0.315022621	4.318051394	15.53728789	0.31929324
KURTOSIS (moment 4)	-0.345336407	42.0473417	84.093711	28.46291164	4.479026691	-1.065396421	20.91890929	389.4303894	-0.511660733
MIN	0.11	4	0	-43.338	1	3	1	-12465.9	0
MAX	290.89	9999	320680.3	2759.284	9	10743	232	229560.7	157.86
QUARTILE 1	80.2	34	877.76	2.86	1	2947	6	77.74728	20.906525
QUARTILE 2	215.91	65	3691.15	19.67	2	4978	14	879.9441	63.754
QUARTILE 3	471.54	572	11995	123.21	3	7971	25.5	3035.413	91.66355
COUNT	3389	3389	3389	3389	3093	3389	1551	3389	1564

MAY2016 PR 3	MONTANT_AUTORISE_FORCAGE_modif	PROBABILITE_DEFAUT	GLOBAL_ASSETS	MT_REVENU_TOTAL	Nb_Doscred	Anciencpt	V11_UNI	V32_UNI	V51_UNI
TESTS STATS									
Average (moment 1)	283.6687514	788.2932417	6073.68921	84.91384192	1.949061662	4819.751432	37.28621908	1390.910367	75.2829515
STD (moment 2)	266.160674	1496.384694	18419.47266	187.6576311	1.085637995	2975.82914	52.76928633	4782.737372	37.72442434
SKEWNESS (moment 3)	0.965972659	4.485497987	8.951186537	8.712184571	1.287761022	0.628656423	4.295684582	7.309306364	0.106496565
KURTOSIS (moment 4)	-0.151551419	23.94825384	110.1251629	140.7635788	1.721764709	-0.614747825	27.97899361	80.35724207	0.989250543
MIN	0.26	4	0	-32.87	1	116	1	-12141.2	0
MAX	992.93	9999	252735.9	3580.18	6	10743	523	60257.51	200
QUARTILE 1	66.34	34	25	2.65	1	2307	12	-155.239	54.0931
QUARTILE 2	193.67	212	597.35	18.22	2	4331	21	235.2302	84.1313
QUARTILE 3	451.86	1145	3513.37	75.54	3	6464	38	1071.573	97.1146
COUNT	873	873	873	873	746	873	566	873	339

	MONTANT_AUTORISE_FORCAGE_modif	PROBABILITE_DEFAUT	GLOBAL_ASSETS	MT_REVENU_TOTAL	Nb_Doscred	Anciencpt	V11_UNI	V32_UNI	V51_UNI
F TEST DISPERSION	0.465535613	4.88274E-14	1.06797E-07	0.482385422	1.21964E-31	0.005570852	5.12978E-13	2.19339E-67	0.005037877
T TEST ESPERANCE	0.045332551	9.20114E-09	9.40428E-09	0.013695777	1.81011E-33	1.87564E-08	3.90387E-07	1.45433E-08	7.6986E-11

Les variables principales retenues sur ce segment de clientèles au seuil de 10% concernent les historiques de comptes, des probabilités des risques de crédit, le nombre des encours de crédit et des engagements de la banque, de l'ancienneté de la relation commerciale, des actifs sous gestion, des revenus des clients…

Les découverts bancaires sont plus acceptables :

• lorsque la probabilité de faillite est faible (les probabilités sont calculées à l'aide de méthodes de régression logistique).

• lorsque le client dispose d'actifs (autres comptes et placements).

• lorsqu'il est un client actif (un client actif est un client rentable au sens du contrôle de gestion).

• lorsqu'il est fortement engagé auprès de la banque. Il est plus difficile de refuser un découvert lorsque l'entreprise doit rembourser ses dettes auprès de la banque.

• lorsque l'ancienneté du client est importante. Il est plus facile d'accepter un dépassement pour un client de longue date.

• lorsque les indicateurs d'exploitation du compte sont bons (V11_UNI – V32_UNI : solde moyen du compte – nombre de jours d'excédent…).

Statistiquement, nous n'avons pas mis en évidence un effet de saisonnalité significatifs au seuil de 5% avec le test du Khi Deux sur les décisions de découverts et ce quel que soit le type de clientèle. Cet aspect sera développé dans un autre article explorant l'impact de l'I.A. sur la « supply chain » financière des entreprises.

2. Les méthodes possibles et justifications

Nous avons préalablement écarté le « deep learning » pour des raisons techniques et de dimensions des variables.

Dans ces démarches positivistes, les décisions binaires supervisées peuvent être appréhendées par de multiples méthodes. Pour l'historique rapide, citons l'analyse discriminante multivariée et les régressions logistiques. Pour s'affranchir des hypothèses statistiques classiques, les méthodes C.A.R.T. « *Classification And Regression Trees* » et M.L. sont actuellement les méthodes les plus utilisées dans les processus d'analyse y compris dans l'I.A.R.D. des assurances qui abandonne progressivement les méthodes G.L.S. pour calculer les primes de risque.

2.1. La méthode C.A.R.T.

L'algorithme commence par choisir la variable explicative, qui grâce à ses modalités, découpe le mieux la population en deux groupes en maximisant la variance inter-groupe. Les deux

groupes créés sont appelés « nœuds ». On renouvelle l'opération jusqu'à ce qu'il n'y ait plus qu'un individu par groupe ou selon un critère d'arrêt à définir, ce qui permet d'obtenir les nœuds finaux appelés feuilles.

La seconde étape propose de minimiser une fonction prenant en compte l'erreur quadratique moyenne et le nombre de feuilles. Cette fonction permet d'optimiser le niveau de complexité de l'arbre de manière à éviter le sur-apprentissage. Le sur-apprentissage serait de faire de chaque cas une feuille. On obtient un arbre optimal.

La méthode C.A.R.T. dépend fortement de l'ordre des variables choisies et des variables choisies pour réaliser le modèle prédictif, D'où l'importance de réaliser correctement la partie sélective des variables. Cette limite peut être rectifiée par des techniques de boosting ou bagging proposées par les Machines Learning.

L'énorme avantage des méthodes C.A.R.T. est de comprendre les décisions grâce à des nœuds simples. On obtient une décision ou un ensemble de décisions en passant successivement par une suite d'inéquations de variables explicatives (rectangles bleus du tableau 2 présenté ci-dessous). Une feuille est correcte lorsque l'on obtient une majorité de types de décisions (dominance de oui ou de non ou rond vert et rouge du tableau 2 présenté).

2.2. Machine Learning M.L.

La méthode M.L. propose un apprentissage global ou ensembliste à partir de la technique du bagging. L'objectif est d'entraîner plusieurs modèles pour pouvoir proposer un modèle final qui combine leurs sorties. Le bagging crée plusieurs sous-ensembles de données d'apprentissage par échantillonnage aléatoire avec remise. Le bagging permet d'améliorer la stabilité et la précision des prédictions par rapport à un modèle obtenu à partir d'un algorithme d'apprentissage. Il aide à réduire l'entropie et à éviter le surapprentissage.

Techniquement on cherche pour chaque scission, non plus la meilleure scission parmi toutes variables explicatives (n), mais la meilleure scission pour p variables explicatives tirées aléatoirement parmi n. Cette double « randomisation » a été introduite par L. BREIMAN [1]. Le nombre d'arbres dans la forêt croît avec le nombre de variables.

Malgré toutes les subtilités de paramétrages, les modèles ne peuvent échapper au sur-apprentissage. Nous reviendrons plus tard sur le sujet dans cet article et dans la mise en place d'un ensemble d'algorithme. Plus les algorithmes sont complexes, plus le risque opérationnel associé est plus important. Plus le processus est complexe, plus la traçabilité des variables explicatives est faible. Ces méthodes, globalement plus complexes que les méthodes statistiques, sont souvent assimilées à des boites noires en raison de la subtilité des algorithmes et de leurs paramétrages. Les M.L. sont confrontées aux logiques d'interprétations et de partage des résultats. Enfin, Il faut cependant reconnaitre que les M.L. sont plus simples à mettre en œuvre que le « *deep learning* » présenté dans le tableau 3 avec des variables qui ne sont pas nécessairement explicatives. Par conséquent, les M.L. doivent être utilisées de manière croisée avec les approches classiques pour affiner les hypothèses et les choix. Le tableau 2 synthétise les approches différentes entre le C.A.R.T. (arborescence décisionnelle avec des inéquations) et le *random forest* (apprentissage ensembliste de plusieurs modèles)

Tableau 2 : C.A.R.T. et Random Forest [2]

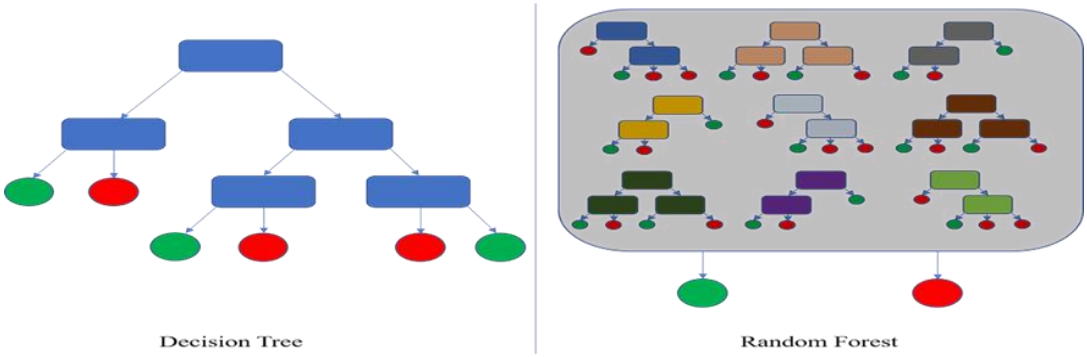
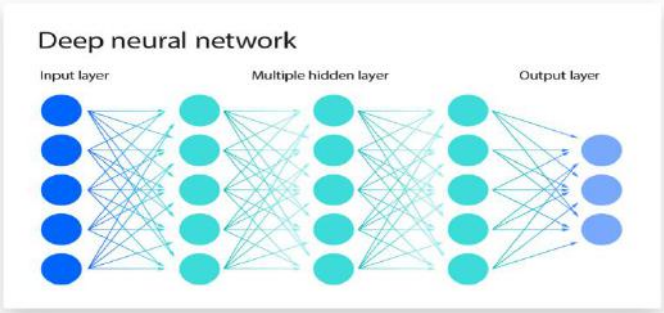


Tableau 3 : deep learning [3]



3. Les choix méthodologiques pour la mise en place d'une M.L. et ses résultats

La création de la base de données a été un processus long et complexe, en raison de la multiplicité des sources d'information et des bases de données.

Un audit technique et analytique a été nécessaire pour identifier les bonnes bases, les bonnes variables et de créer une base réplicable avec les données historiques nécessaires.

Le choix technologique a été aussi un sujet au sein de la direction de la banque. Pour des raisons de coût, il a été choisi des méthodes standards disponibles dans des librairies R ou

[2] https://www.pericles-group.com/
Machine Learning : Du GLM à l'arbre de CART en passant par le Random Forest
A Guide to Random Forest in Machine Learning Ejable novembre 2023

[3] https://www.ibm.com/think/topics/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks

Python. Les méthodes standard pour la M.L. sont la méthode S.V.M. « *support vector machine* » et K.N.N. « *K Nearest Neighbour* ».

De nombreuses études montrent les différences de performances entre les différentes techniques de M.L.. Citons l'article de Danilo Bzdok, Martin Krzywinski, Naomi Altman [3].

3.1. K.N.N.

K.N.N. est un algorithme d'apprentissage automatique supervisé simple et très efficace. Il appartient à la famille des algorithmes non paramétriques en se basant sur la similitude des points de données d'entrée. K.N.N. fait essentiellement des prédictions basées sur la similitude des points de données dans l'espace d'échantillonnage. Les performances de K.N.N. sont essentiellement basées sur le choix de K.
K.N.N. fonctionne en mémorisant l'ensemble des données d'entraînement. Lorsqu'un nouveau point de données est donné pour la prédiction, K.N.N. examine les points de données les plus proches dans l'ensemble d'entraînement en fonction d'une métrique de distance spécifiée (généralement la distance euclidienne). Pour la classification, il attribue la classe majoritaire parmi les k voisins les plus proches au nouveau point de données. Pour la régression, elle prédit la moyenne ou la moyenne pondérée des valeurs cibles des k voisins les plus proches.

- Avantages de *K Nearest Neighbour*(K.N.N.)

Sa mise en œuvre est simple. K.N.N. est facile à comprendre et à mettre en œuvre, ce qui le rend adapté au prototypage rapide.
K.N.N. est un algorithme d'apprentissage qui ne nécessite pas de phase d'entraînement. Le modèle est construit pendant la phase de prédiction.

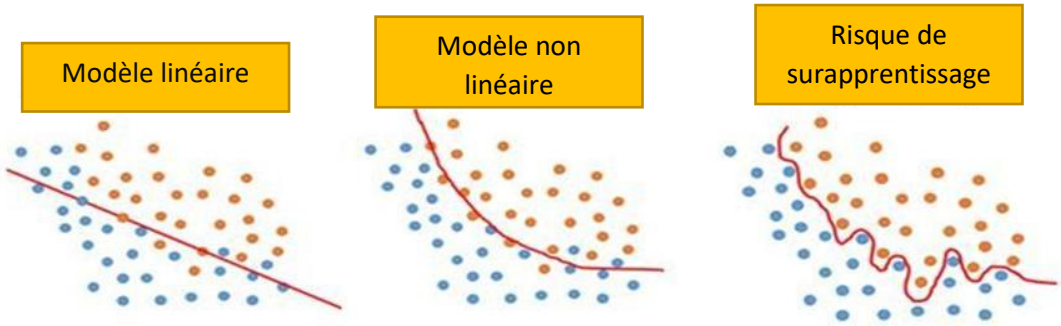
- Inconvénients de K.N.N.

L'inconvénient majeure est la complexité de calcul. Cette dernière augmente avec la taille de l'ensemble de données. Le calcul nécessaire pour trouver les voisins les plus proches augmente, ce qui entraîne des coûts de calcul plus élevés. Cette méthode est également sensible aux valeurs aberrantes, en affectant de manière significative les distances entre les points et, par conséquent, les prédictions.

3.2. La méthode S.V.M.

S.V.M. trouve l'hyperplan optimal qui maximise la marge entre les points de données les plus proches des classes opposées. La marge est la distance entre l'hyperplan et les échantillons les plus proches. Ces derniers sont appelés vecteurs supports. L'algorithme S.V.M. est largement utilisé dans le M.L., car il peut effectuer des classification linéaires et non linéaires. Lorsque les données ne sont pas linéairement séparables, des fonctions de noyau sont proposées pour transformer les données dans un espace à plus haute dimension afin de permettre une séparation linéaire. Cette application des fonctions du noyau est connue sous le nom de « astuce du noyau » ou (kernel trick). La méthode S.V.M. permet des classifications linéaires souvent peu adaptées et des classifications non linéaires (polynomiales…) comme présenté dans le tableau 4.

Tableau 4



Pour résumer nous pouvons avancer une matrice forces et faiblesses de la méthode.

- Avantages de la machine à vecteurs de support (S.V.M.) :

Efficacité dans les espaces de grande dimension. Le S.V.M. fonctionne bien dans les espaces de grande dimension, ce qui le rend adapté aux tâches avec un grand nombre de fonctionnalités.
Elle est robuste au surajustement. En effet le S.V.M. dispose de paramètres de régularisation qui aident à éviter le surapprentissage tout en proposant des modèles non linéaires pour séparer les espaces de classification.
C'est une méthode d'optimisation globale qui nécessite un problème convexe. Le cas échéant, cela garantit que la solution trouvée est l'optimum global. Efficace dans les données non linéaires : avec l'utilisation de fonctions de noyau, S.V.M. peut gérer des relations non linéaires entre les entités.

- Inconvénients de la machine à vecteurs de support (S.V.M.) :

L'optimum global entraine une complexité des calculs lorsque l'on a un grand ensemble de données. Elle est gourmande en mémoire : en particulier lorsqu'il s'agit de grands ensembles de données, car l'algorithme doit stocker tous les vecteurs de support. La méthode est sensibilité au bruit : la S.V.M. est sensible aux données bruyantes, et les valeurs aberrantes dans l'ensemble d'entraînement peuvent avoir un impact significatif sur les performances.
La sélection d'un noyau approprié et son ajustement par des paramètres peuvent être difficiles, et les performances du modèle S.V.M. sont sensibles à ces choix.

3.3. L'arbitrage méthodologique

Nous avons choisi une combinaison S.V.M. et C.A.R.T. Ces méthodes hybrides se retrouvent dans la littérature scientifique [9].

Les méthodes comme présentées sont toutes imparfaites. L'inconvénient des M.L. est globalement d'échapper partiellement aux processus explicatifs de la décision et reste sensible à la qualité des données d'apprentissage.

Dans notre recherche et mise en place, nous avons à ce stade les données et les variables essentielles grâce aux tests d'hypothèse. Nous avons utilisé C.A.R.T. pour comprendre le processus décisionnel et identifier des anomalies dans la base de données initiale. En effet des réponses positives de découverts dans des configurations de réponse négatives ont soulevé de grandes interrogations. L'identifications des cas a permis de mieux comprendre les cas générant d'importants bruits. Nous avons trouvé concrètement des cas ou la décision

ne peut pas être comprise sans la prise en compte de variables manquantes. Un exemple que l'on peut citer sont des cas d'acceptation de découvert alors que le client a tous les attributs d'un mauvais client (manque d'avoir – risque – faible rentabilité pour la banque). Apres une étude du dossier client concerné, nous avons constaté que la banque a autorisé le découvert grâce à une garantie juridique de qualité provenant de la maison mère de l'entreprise. Faute d'incorporer cette information, la décision est incompréhensible voire dangereuse dans la mise en place. Il existe deux solutions. La première est d'intégrer cette variable manquante (garantie et qualité de cette dernière). La deuxième solution est d'élaguer dans le processus d'apprentissage les informations identifiées comme bruitées. La banque a retenu cette deuxième solution faute de pouvoir construire de manière fiable cette variable informatique. Ces données identifiées et supprimées de la base d'apprentissage ont représenté 3.5% de la base de données d'origine.

Nous avons ensuite utilisé la méthode SVM sur les données ou décisions épurées, choix fondé sur la qualité de sa méthode d'optimisation globale.

De plus, S.V.M. est adaptée à notre problème de deux classes distinctes (les acceptations et les refus de dépassement de soldes débiteurs).

3.4 Les résultats

Il existe de multiples critères pour évaluer la qualité statistique du modèle M.L. proposé. Le plus utilisé est naturellement la matrice de confusion qui permet d'identifier les bonnes réponses et les autres, ou autrement dit les risques d'erreurs primaires et secondaires. Nous rappelons que notre base de données comporte 432 966 décisions avec 373 389 autorisations de découverts et 59 577 refus réparties entre 7 segments de clientèles. Nous avons évidemment réalisé 7 modèles qui se sont révélés différents tant par leur composition, que par leur performance.

Dans notre étude, vous avons 4 cas:

Nous avons les autorisations de découvert autorisés correctement modélisées. (vrai positif)

Les refus de découverts correctement modélisés (vrai négatif)

Les autorisations de découverts non acceptées par le modèle (faux positif)

Les refus de découverts acceptés par les modèles (faux négatif)

L'information pour l'ensemble des segments est présentée dans le tableau 5.

Il est évident que les cas des faux positifs et négatifs posent des problèmes d'acceptabilité des managers de la banques. Les faux négatifs autorisent à tort des acceptations de découverts. Cette pour cette raison que la banque limite les seuils d'acceptation de découvert automatisés. Les faux positifs posent aussi problème car le compte peut être bloqué. Ce dernier cas n'est pas problématique au sens ou la banque conserve la décision finale de ce blocage en déclenchant manuellement une cessation de paiement. Globalement les performances des modèles se sont révélées de bonne qualité avec le critère de la performance globale « accuracy » (Vrai positif + vrai négatif) / total. Il est à noter que les faux positifs sont faibles avec 2 538 cas sur un total de 59 577 non acceptation (4.26% de cas).

(316 772 + 57 039)/432 966 = 87.10 %

Tableau 5 : résultats – matrice de confusion

	Modélisation oui	Modélisation non
Acceptation	316 772 Vrai positif	56 617 Faux négatif
Refus	2 538 Faux positif	57 039 vrai négatif

Tableau 6 : performance par segment

Segment clientèle	S1	S2	S3	S4	S5	S6	S7
« accuracy »	92.40	96.78	79.73	27.61	88.13	57.57	78.36

Nous avons constaté que le modèle ne fonctionne pas pour faire face aux entropies importantes dans deux segments de clientèle (S4 et S6). Ces deux segments concernent la clientèle « retail » haut de gamme qui est la moins nombreuses en termes d'effectif par rapport aux entreprises et aux clients personnes physiques classiques. L'entropie de S4 et S6 est liée à la complexité et à la singularité des situations de la clientèle haut de gamme. Le tableau 6 montre la performance globale pour les 7 segments.

Les résultats montrent que les automatisations ont été effectives sur les segments S1 – S2 – S3 - S5 et S7 avec une performance globale pondérée de plus de 90%.

La banque a cependant limité les automatisations automatisées sur des montants limités de 500 à 1000 euros en fonction du segment de clientèle. Cette étude montre les avantages et les limites de cette automatisation. Les avantages sont clairement une automatisation maitrisée sur certain segment de clientèle avec des gains en temps considérables pour les employés de la banque. L'IA est une source de gain de productivité mais doit rester un processus maîtrisé.

L'IA engendre aussi des couts de développement et de maintenance qui restent non négligeables.

Les sources d'amélioration restent dans le perfectionnement des bases de données et les avancées méthodologiques que l'on peut réaliser dans le domaine des Machines Learning. Ici, le problème est l'asymétrie des décisions entre les autorisations et les refus. Cette asymétrie peut limiter les performances des méthodes classiques de type S.V.M.

4. Une nouvelle méthode créée spécifiquement pour répondre à la problématique de la méthodologie : l'absence de symétrie dans le nombre de décisions

Les données dans notre domaine d'études sont évidemment non symétriques entre les autorisations de découverts beaucoup plus nombreuses que les refus. Ceci est un challenge pour obtenir des algorithmes robustes et fiables. L'absence de fiabilité et de robustesse peuvent entrainer des conséquences légales lourdes pour les banques.

Pour faire face à ces asymétries de décisions, les approches les plus récentes proposent soit des combinaisons méthodologiques comme présenté antérieurement (C.A.R.T. – M.L. – régression logistique…) [7, 12, 13] soit des avancées méthodologiques sur les M.L. [15, 17].

4.1 présentations de la fonction DC programming and D.C.A.

La méthode SVM propose une optimisation globale. Nous avons réalisé dans le cadre d'un programme doctoral des recherches [15, 17] une adaptation de la méthode en proposant la méthode DC *programming* et D.C.A.. Cette méthode a prouvé dans de nombreux problèmes d'optimisation complexe son avantage en termes de résultat et de temps de calcul [14]. La complexité des méthodes d'optimum est souvent le caractère non convexe du problème. La Méthode DCA permet de résoudre le problème de convexité en transformant la fonction initiale en deux fonctions convexes différenciées.

La formulation est la suivante

$$\inf\{f(w) = G(w) - H(w) : w \in \mathrm{R}^{\mathrm{p}}\}, (\mathrm{P}_{\mathrm{dc}})$$

Dans le contexte donné, nous avons les fonctions convexes G et H ∈ Γ0 (Rp), qui sont l'ensemble des fonctions convexes semi-continues inférieures propres d'un ensemble Rp à RU {+∞}. Ces fonctions sont appelées fonctions CC, où G−H représente une décomposition DC de la fonction f, G et H étant les composantes DC. Une contrainte convexe w ∈ C peut être incorporée dans la fonction objective de (Pdc).

4.2. Echantillonnage pondéré sensible aux coûts basés sur D.C.A. (CSB – DCA)

L'objectif du M.L. est d'entraîner plusieurs modèles ou combinaison de variables pour pouvoir proposer un modèle final qui combine leurs sorties. Le bagging crée plusieurs sous-ensembles de données d'apprentissage par échantillonnage aléatoire avec remise.

Lorsque l'échantillon est déséquilibré la méthode peut être perturbée par la prépondérance d'une décision. Notre étude comporte effectivement beaucoup plus de réponses favorables que de réponses négatives de découvert. L'apprentissage sensible aux couts permet de corrigé cette asymétrie. La méthode consiste à affecter un poids plus important à la classe minoritaire. En pratique, le modèle considère que le fait de bien classer une décision de la classe minoritaire (ici le refus) est plus important que de bien classer une décision de la classe majoritaire (l'acceptation). C'est pour cela que la technique S.V.M. et C.A.R.T. préalablement mis en place avait comme défaut de créer un déséquilibre important entre les faux positifs et les faux négatifs.

Comme présenter dans le tableau 2, chaque réponse formalisée est issue d'une agrégation pondérée de différents modèles. Nous pouvons la formaliser de la façon suivante :

$$f(x) = \sum_{j=1}^{n} w_j f_j\ (x)$$

Ou *f(x)* est l'agrégation des modèles et *w* le poids de chacun.

La détermination des poids appropriés est cruciale pour une performance efficace du modèle. Dans le schéma d'ensachage standard, chaque modèle est pondéré de manière

identique. Breiman [2] propose de surpondérer logiquement les modèles proposant les choix les plus majoritaires.

Étant donné un ensemble de données d'apprentissage$\{(x_i, y_i)\}$ à m où m est le nombre d'échantillons de l'ensemble de données d'apprentissage. Chaque échantillon est associé à une étiquette yi (1 ou -1). La sortie du modèle agrégé pour le i-ème point de données est notée $\hat{y}_i$. L'objectif est de minimiser l'erreur de prédiction, ce qui donne lieu à un problème d'optimisation qui sera ici solutionné avec la méthode C.S.B. – D.C.A. « cost sensitive based D.C.A. » :

$$min\ k(w)_{w\epsilon R^n} = \frac{1}{m}\sum_{i=1}^{m} l(y_i, \hat{y}_i) = \frac{1}{m}\sum_{i=1}^{m} l(y_i, \sum_{j=1}^{n} w_j\ f_j\ (x_i))$$

$$w_j \geq 0, \forall j = 1, \ldots, n$$

où l est une perte permettant de mesurer l'écart entre les valeurs prédites et les véritables valeurs.

4.3. Les résultats comparatifs avec le segment de clientèle S2

Une comparaison de performance a été réalisée entre les deux approches C.A.R.T. – S.V.M. et C.S.B. – D.C.A. avec les critères habituels, performance globale « accuracy », Fscore, Gmean et AUC [4] sur le segment clientèle S2.

Tableau 7 : comparaison des performances pour le segment clientèle S2

	Accuracy	Fscore	G mean	AUC
CART - SVM	97.20	98.28	98.57	98.56
CSB - DCA	99.98	99.75	99.78	99.78

Le modèle C.S.B. – D.C.A. montre une légère amélioration des critères de convergence et une supériorité par rapport à la combinaison C.A.R.T. – S.V.M. Cette méthode adaptée aux asymétries des réponses permet notamment de réduire les faux positifs.

5. Conclusion et recherche ultérieure

[4] Fscore = vrai positif / (vrai positif + faux positif).
L'AUC est une mesure de la capacité du modèle à faire la distinction entre Classes positives et négatives. Le ROC (Receiver Operating Characteristics) courbe est une représentation graphique des performances du modèle, traçant le véritable taux de positifs (TPR) par rapport au taux de faux positifs (FPR) pour différents seuils Paramètres. Pour calculer l'AUC dans Matlab, nous utilisons la règle trapézoïdale dans la fonction « Trapz ». La valeur de l'AUC varie de 0 à 1, où un score plus élevé indique un meilleur modèle.

G-mean est la moyenne géométrique des taux de réponse vrai positif

$$\sqrt{\frac{vrai\ positif}{vrai\ positif\ +\ faux\ négatif} \times \frac{vrai\ positif}{vrai\ positf\ +\ faux\ positif}}$$

Cette mise en place d'une IA dans une banque montre les différentes dimensions qui s'imposent à une recherche dans ce domaine. La première dimension est économique et managériale en modifiant profondément l'organisation du travail au sein des banques et plus généralement au sein des entreprises. L'IA est une source de productivité et de réaffectation des ressources. Le deuxième enjeu est de comprendre les systèmes d'information très complexes qui doivent servir de base d'apprentissage au M.L.. Cette complexité nous amène à favoriser les méthodologies hybrides, ici en mixant la méthode C.A.R.T. et S.V.M. pour améliorer et comprendre les bases d'apprentissage. Cette approche ne renonce pas pour autant aux méthodes d'analyse des données en amont pour améliorer la compréhension des processus de décision.

La mise en place de cette approche est globalement satisfaisante car elle a permis d'automatiser la plupart des décisions de dépassement des découverts bancaires mise à part des segments clientèles haut de gamme traditionnellement caractérisés par une très forte entropie.

Le troisième enjeu est technologique. La banque a opté pour des librairie open source de type Python pour installer les algorithmes pour des raisons de coût et de comparabilité. Il n'en demeure pas moins que des améliorations méthodologiques existent en M.L. lorsque les réponses à modéliser sont asymétriques. Notre proposition C.S.B. – D.C.A. apporte une amélioration de la qualité de la convergence avec la réalité observée. Cette méthode sera testée sur des segments de clientèle plus étendus et sur sa capacité à résister à l'introduction dans la base de données de variables aléatoires.

Bibliographie

1. Breiman, L. : random forests Published: October 2001 Volume 45, pages 5–32, (2001)
2. Breiman, L. : Bagging predictors. M.L. 24(2), 123–140 (1996)
3. Bzdok, D., Krzywinski, M., Altman, N. : M.L.: Supervised methods, SVM and kNN. Nature Methods, pp.1-6. (2018)
4. Chang, Y.C., Chang, K.H., Wu, G.J.: Application of extreme gradient boosting, trees in the construction of credit risk assessment models for financial institutions.
   Published in Applied Soft Computing 1 December (2018)
5. Damel, P. : Les produits structurés bancaires et le Contrôle de Gestion bancaire : une approche comparative utilisant les taux de marché de référence, communication avec publication dans les actes du colloque, 22ième congrès AFC (Mai 2001)
6. Damel, P. : L'apport des méthodes de « replicating portfolio » ou portefeuille répliqué en A.L.M. : méthode contrat par contrat ou par la valeur optimale », Banque et Marchés Mars-Avril (2001)
7. Kim, M.J., Kang, D.K., Kim, H.B.: Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction.
   Expert Systems with Applications 42(3), 1074–1082 (2015)
8. Hafez, IY. : A systematic review of AI-enhanced techniques in credit card fraud detection Journal of Big Data, SpringerOpen. (2025)
9. Komal Goyal, Megha Garg Shruti Malik Adoption of artificial intelligence-based credit risk assessment and fraud detection in the banking services: a hybrid approach (SEM-ANN) Future Business Journal volume 11, Article number: 44 (2025)
10. Guan, C. : Responsible Credit Risk Assessment with M.L. and Knowledge Acquisition Springer article "Human-Centric Intelligent Systems » SpringerLink (2023)
11. Lennart, H., V., Damásio, B. : M.L. in banking risk management: Mapping decade of evolution. International Journal of Information Management Data Insights
    Volume 5, Issue 1, June 2025, 100324. (2025)
12. Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M.S., Zeineddine, H.: An experimental study with imbalanced classification approaches for credit card fraud
    detection. IEEE Access 7, 93010–93022 (2019)
13. Veganzones, D., S´everin, E.: An investigation of bankruptcy prediction in imbalanced datasets. Decision Support Systems 112, 111–124 (2018)
14. Pham Dinh, T., Le Thi, H.A.: Convex analysis approach to dc programming: theory, algorithms and applications. Acta Math. Vietnam 22(1), 289–355 (1997)
15. Pham Van Tuan (co-supervision Hoai An Lethi et Damel Pascal). "New M.L. techniques in financial decision making". Thesis defended in November 2023, at University of Lorraine (2023)

16. Roy, J.K. : M.L. and Artificial Intelligence Method for FinTech Credit Scoring and Risk Management : A Systematic Literature Review. International Journal of Business Analytics Volume 11 • Issue 1 • January-December (2024)
17. Van Tuan Pham, Hoai An Le Thi, Hoang Phuc Hau Luu, Damel Pascal: DCA-Based Weighted Bagging: A New Ensemble Learning Approach. Published ACIIDS (2) (2023)
18. Habib Zouaoui, H. : Credit card fraud detection and risk management strategies: A deep learning-based approach for EU banks. Research Papers in Economics and Finance Vol. 9 No. 1 (2025)