InData

La lettre connectée à l'actualité digitale

AVRIL 2023

# Machine Learning : Une nécessité pour le secteur bancaire

LENDYS
Pôle Transformation
Finance & Digital

## Introduction

Les développements technologiques de ces deux dernières décennies comme le Big Data et autres technologies de l'Intelligence Artificielle (IA) ont engendré des transformations majeures sur nos comportements. L'entrée dans une ère de plus en plus digitalisée a décuplé la masse de données. Leur exploitation à une échelle sans précédent conjuguée aux progrès réalisés en matière de stockage et de calcul, bouleverse divers secteurs, le financier en particulier. Dans ce contexte, le Machine Learning (ML), une discipline de la science des données, est en pleine expansion. Cet article s'adresse à un public non-initié souhaitant explorer les possibilités d'usages du ML dans le secteur bancaire.

## Nouvelles technologies, nouvelles attentes, nouveaux acteurs

Vous y conviendrez, avec la volumétrie de données, les combinaisons sont multiples et le client a généralement recours à des expertises variées pour l'accompagner dans ses choix. En effet, sous l'impulsion d'une digitalisation accrue, le client s'attend aujourd'hui à pouvoir régler ses opérations à tout moment en ligne. Autrement dit, le nouveau rapport à la donnée (accessibilité, réactivité, simplicité) a modulé significativement l'offre de services dans une logique davantage orientée client que produit.

Si les fondamentaux du métier demeurent, c'est bien aux nouvelles attentes des clients désormais fortement digitalisés qu'il s'agit de répondre : des offres de services simples d'accès, accessibles partout et aux parcours « fluides ». Nous pouvons brièvement citer les exemples de banques en ligne, néo-banques[1] ou encore des fintechs.

Sur un tout autre plan, les techniques de traitement des signaux faibles (analyses de tendances, identification de transactions suspectes, suggestion personnalisée ou encore l'identification de crédit risqué) reposent sur une variété et une masse de données fournissant aux algorithmes la capacité à créer des liens. Pensez que le ML, dont nous avons abordé le principe dans notre précédent article[2] prend ici tout son sens.

*Interrogés sur l'émergence de concurrents et sur leur propre positionnement concurrentiel, les établissements citent en premier lieu les concurrents susceptibles de leur ravir la relation client : les acteurs proposant des services bancaires, ou seulement de paiement, de manière entièrement numérique, en ligne ; c'est-à-dire les « néo-banques ».*

**Analyse et synthèse n°131**
**ACPR-Banque de France**

[1]Une banque digitale, accessible uniquement via une application mobile. En réalité, une néo-banque n'est pas une banque mais généralement un établissement de paiement qui s'appuie sur les innovations technologiques. Pour plus de détails voir : acpr.banque-france_regle_usage_terme_neobanque.pdf

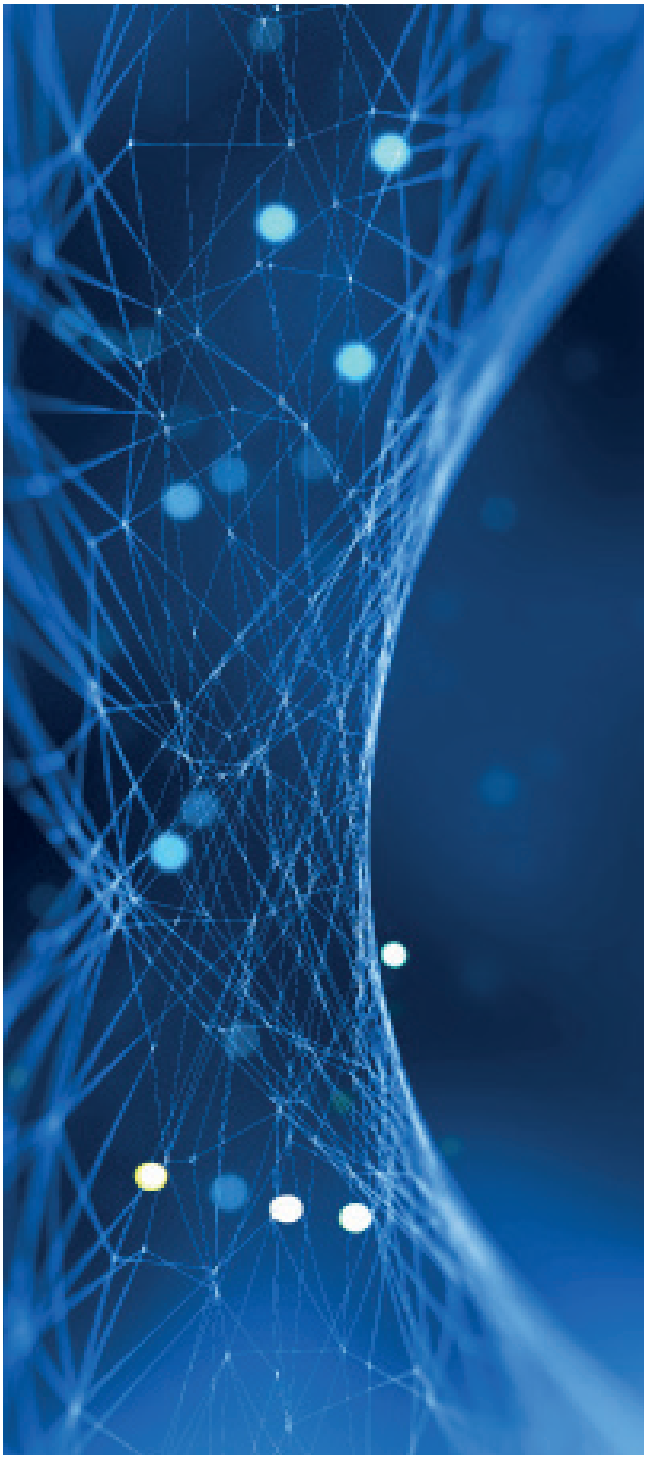
[2]Digitalisation de la fonction finance : Chronique d'une transformation annoncée.
Lien : article digitalisation de la fonction finance

## Le ML : ce que vous devez retenir

Il est toujours utile de rappeler les concepts clés qui sous-tendent cette discipline en constante évolution. Ainsi, le Machine Learning (ML), aussi appelé **« apprentissage automatique »** est une technique d'apprentissage dérivée de l'intelligence artificielle. Une technologie qui confère à un programme la capacité de faire évoluer la performance de ses résultats au fur et à mesure qu'il intègre de nouvelles données. C'est-à-dire progressivement plus précis à mesure qu'ils s'exécutent. Basée sur la puissance du ML et des statistiques, l'analyse prédictive quant à elle, permet d'identifier des tendances dans les comportements passés en vue d'anticiper les besoins ou évènements futurs. Autrement dit, l'utilisation conjointe des modèles statistiques et du ML sur de grands volumes de données délivrent l'aspect « prédictif » permettant aux outils d'analyse prédictive d'aller au-delà de la simple corrélation.

> ***« L'usage accru et responsable des données, notamment avec l'intelligence artificielle, est un levier stratégique*** *sur lequel la Société Générale entend continuer de capitaliser pour adapter et améliorer ses modèles de distribution et de production existants, mais aussi inventer et faire grandir de nouveaux business modèles alternatifs. Devenir une banque 'data-driven', avec une utilisation responsable des données, est au cœur de notre agenda stratégique. »*
>
> **Déclaration de Fréderic Oudéa,**
> Directeur Général de la Société Générale,
> communiqué de presse, Paris 23/11/2021.

**Après ce rapide rappel, faisons un focus sur des exemples d'applications bancaires.**



## Potentiel révolutionnaire du ML en banque

De l'optimisation des relations avec les clients aux capacités inédites de traitement de la data, le rôle grandissant de l'analyse prédictive dans le secteur bancaire devient un avantage compétitif. En effet, sous l'impulsion des contraintes règlementaires[3] et en raison de la diversité très importante des données historiquement accumulées, la banque dispose d'un atout majeur.

Nous l'aurons compris, la donnée est un actif stratégique pour la banque et l'enjeu est de taille :

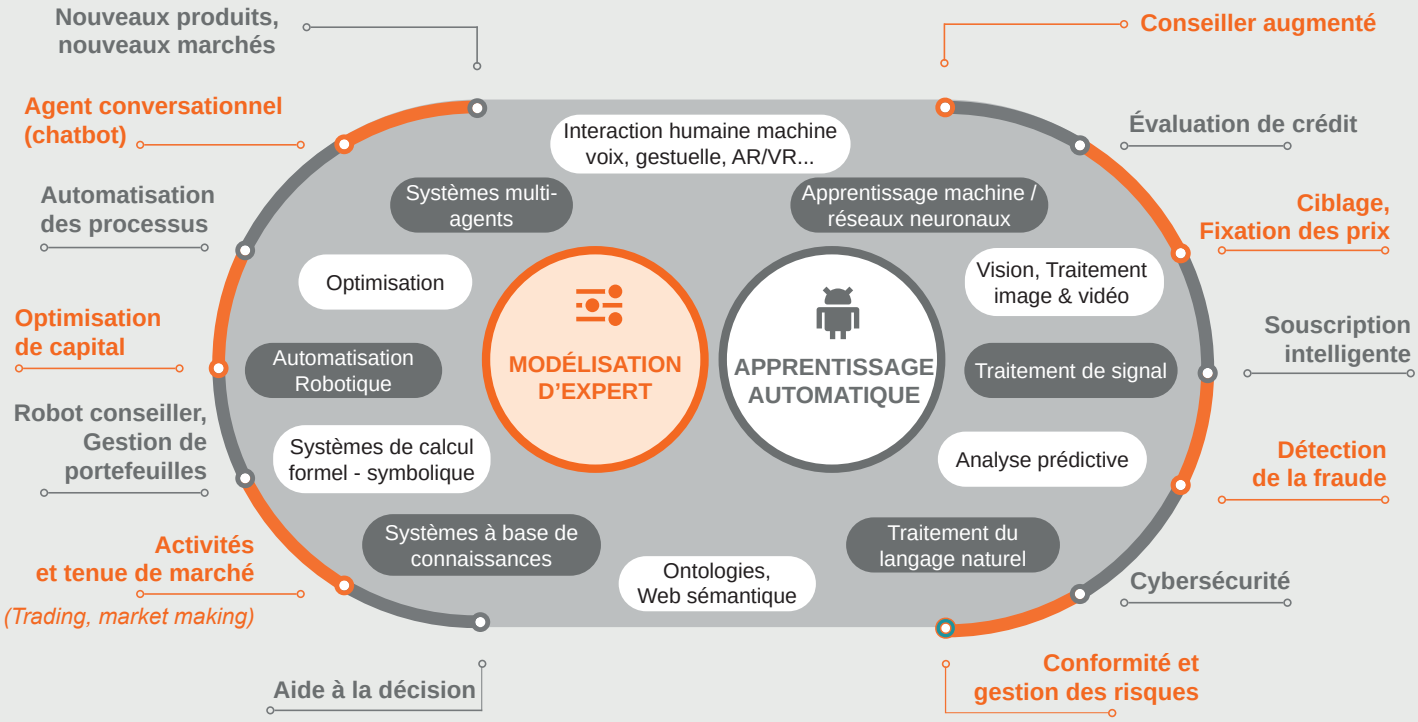
► Exploiter les données pour conserver l'avance par rapport aux concurrents.

Comme vous pouvez l'observer, les possibilités d'applications dans le secteur bancaire sont multiples et variées. Dans cet article, nous avons fait le choix d'explorer quelques cas d'utilisations standards du ML :

► L'analyse prédictive du risque de crédit et la détection de fraude. L'objectif est double.

Dans un premier temps, nous aborderons volontairement, dans un style de vulgarisation, le mécanisme d'apprentissage sur lequel repose le ML. Cette partie est essentielle à la compréhension de l'intérêt du ML dans l'analyse Prédictive. Dans un deuxième temps, nous illustrerons, schémas et exemples à l'appui, les domaines d'applications précitées. Le but étant de mieux apprécier le champ des possibilités du ML. Des définitions viendront ponctuer certains concepts clés du ML pour non-initiés.

Différentes technologies d'IA applicables au domaine financier




*Source Intelligence artificielle, blockchain et technologies quantiques au service de la finance de demain, Finance Innovation, RB Edition, avril 2019*

[3] https://acpr.banque-france.fr/sites/default/files/medias/documents/201902_acpr_ld_identification_verification_et_connaissance_de_la_clientele.pdf

## Comment la machine apprend à apprendre ?

La réponse à cette question se trouve dans la définition d'une programmation. En effet, habituellement nous élaborons un calcul (un programme) que l'ordinateur va résoudre. Mais que se passe-t'il quand nous ignorons le programme ? Par exemple comment reconnaitre l'étiquette d'une image parmi tant d'autres images ? Comment traduire instantanément des textes et discours d'une langue à une autre ?

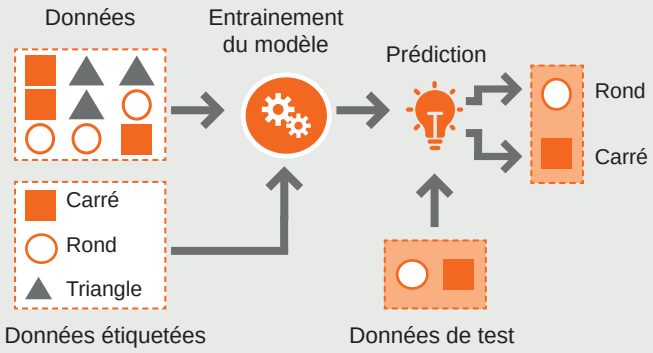
Pourtant des solutions de plus en plus significatives existent et les exemples ne manquent pas. Google, Netflix, Amazon, Spotify, Tesla, Facebook, Microsoft, IBM et tant d'autres utilisent le ML. Du classement automatique des images, en passant par la détection automatique de contenus jusqu'au projet futuriste de la voiture autonome, les possibilités sont aussi vastes que les idées que nous aurons. C'est de là que vient la notion d'apprentissage ou Learning.

Faute de programmation en amont, tout comme nous pouvons apprendre à travers des exemples, la machine va être entrainée à apprendre en lui fournissant en continu des données et donc des exemples (Dataset) comme input. Ce processus inspiré du fonctionnement du cerveau humain peut par exemple être supervisé ou non supervisé.

> *« Machine Learning is the science of getting computers to learn without being explicitly programmed. »*
>
> **Arthur Samuel, 1959.**

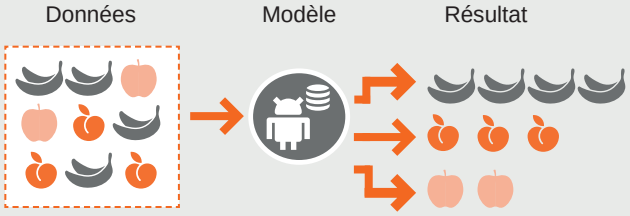
### Apprentissage supervisé

Dans le cas de l'apprentissage supervisé, les données d'entrainement sont étiquetées ou labelisées (carré, rond, triangle) :




### Apprentissage non supervisé

Dans le cas de l'apprentissage non supervisé, les données d'entrainement ne sont pas étiquetées ou labelisées. L'algorithme va chercher à classer les données par caractéristiques communes :




**REMARQUE :**

Nous nous trouvons ainsi face à deux principales typologies de fonctionnement en termes d'automatisation et d'apprentissage aux finalités différentes. En résumé, là où le raisonnement humain trouve sa limite, le Machine Learning aura l'avantage de faire ressortir des informations de manière automatique à partir d'un gros volume de données. L'algorithme utilisé aura la faculté d'assimiler des quantités de données toujours plus grandes afin de générer des résultats et détecter des tendances (pattern). Les résultats seront d'autant plus précis et pertinents que le volume de données sera massif.

# L'application du ML au risque de crédit : Notation (Scoring)

Traditionnellement, le secteur des services financiers utilise des méthodes statistiques dans l'estimation des risques d'insolvabilité. Dans ce cas, les banques conduisent une analyse statistique[4] pour noter l'emprunteur en fonction de sa solvabilité. Le recours au ML dans ce cas représente un atout complémentaire majeur afin de pouvoir identifier au mieux le risque de crédit associé en se basant sur au moins deux aspects supplémentaires.

Tout d'abord, l'utilisation de techniques d'apprentissage automatique permettent de réaliser des prévisions en prenant en compte les relations non linéaires entre les variables explicatives et le risque de crédit individuel, tout en évitant les risques de « surapprentissage »[5]. Ensuite, grâce à la numérisation de nos habitudes et à l'intelligence artificielle, nous pouvons accéder à de nouvelles formes de données. Par exemple des informations provenant de réseaux sociaux tels que LinkedIn, Twitter ou Facebook constituent une autre source d'informations utilisée par les algorithmes d'IA. Afin de mieux appréhender nos propos, prenons l'exemple d'un Dataset issu de Kaggle[6] et illustrons l'intérêt du ML sur une problématique de Credit Scoring.

## 01 BUSINESS PROBLEM

Le groupe « Home Credit » est une agence de prêt de 24 ans (fondée en 1997) qui fournit des prêts à la consommation à ses clients et a des opérations dans 9 pays au total. Ils ont pénétré le marché indien en 2012 et ont servi plus de 10 millions de consommateurs dans le pays. Afin de motiver les ingénieurs ML à construire des modèles efficaces, ils ont conçu un concours Kaggle pour la même tâche. Leur devise est de donner des pouvoirs aux clients non desservis (ceux qui ont peu ou pas d'historique de crédit) en leur permettant d'emprunter facilement et en toute sécurité, en ligne et hors ligne.

## 02 DATASET

Les jeux de données contiennent de nombreuses caractéristiques telles que le type de prêt, le genre, la profession ainsi que le revenu du demandeur, s'il/elle possède une voiture ou un bien immobilier, pour ne citer que quelques exemples. Il comprend également l'historique de crédit passé du demandeur.

## 03 OBJECTIF

L'objectif est d'identifier les profils potentiellement défaillants sur la base du Dataset et du Feature Engineering (technique consistant à créer de nouvelles variables ou « features ») à partir des données existantes afin de mieux représenter les informations et améliorer les performances des modèles de Machine Learning : transformation de variables existantes, la combinaison de plusieurs variables en une seule, la création de variables indicatrices, etc...

## 04 CONTRAINTES

Être en mesure de générer des estimations de probabilité, de la capacité d'un demandeur à rembourser ou non un prêt, plutôt que de strictement classer le demandeur comme l'un ou l'autre (Par exemple : si la probabilité est de 0,5 pour la capacité d'un demandeur et de 0,9 dans l'autre cas, nous pouvons très bien conclure que nous sommes beaucoup plus sûrs de la capacité lorsque la valeur est de 0,9 plutôt que lorsque la valeur est de 0,5).

## 05 FORMULATION DU PROBLÈME

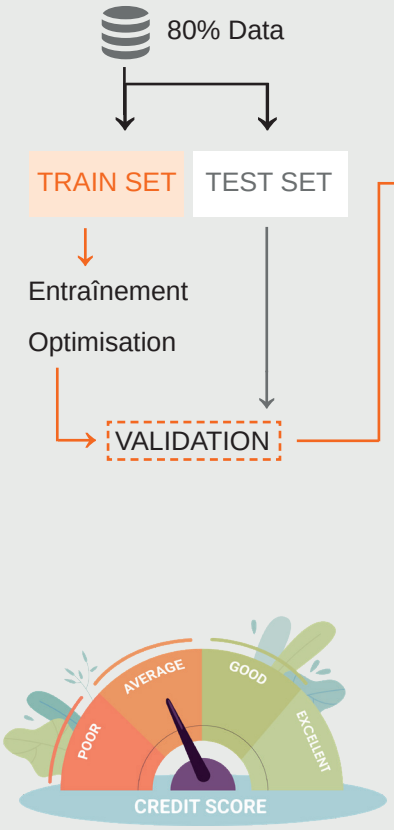
Le jeu de données que nous avons comprend à la fois les caractéristiques et l'étiquette de classe (target), et nous devons prédire les étiquettes de classe correspondantes (0 : Non Défaillant ou 1 : Défaillant). Cela signifie essentiellement que c'est un problème de classification binaire d'apprentissage supervisé.

[4]Voir Hurlin, C. et Pérignon, C. (2019) pour une synthèse.

[5]Le terme anglais est « overfitting ». Quand un modèle a trop appris les particularités de chacun des exemples fournis en exemple. Il présente alors un taux de succès très important sur les données d'entraînement (pouvant atteindre jusqu'à 100%), au détriment de ses performances générales réelles.

[6]Kaggle est une plateforme web organisant des compétitions en science des données : https://www.kaggle.com/competitions/home-credit-default-risk

Le jeu de données prêt[7], il est séparé en deux : un training set et un testing set. Le training set aura pour objectif d'entrainer les modèles sur ces données ensuite nous pourrons faire une comparaison des résultats obtenus avec les testing set. Rappelons que nous sommes dans un cas de classification binaire où le problème est associé à la catégorisation des prêts des clients potentiels en bonne ou en mauvaise classe. Pour une nouvelle demande de prêt, le modèle génèrera un score (une probabilité). Si ce score est inférieur au seuil optimal que nous aurons décidé (par exemple 50%), le prêt est approuvé ; sinon il est rejeté.




Nous déciderons également de la métrique de performance que nous allons utiliser compte tenu des contraintes évoquées précédemment. Par exemple, une matrice de confusion peut être construite dans les scénarios de classification binaire et multi-classe, et est une représentation visuelle très utile pour avoir une vue d'ensemble de toutes les prédictions faites sur une classe particulière par un modèle. Elle montre précisément les classifications correctes ainsi que les mauvaises classifications faites par un modèle. Une matrice de confusion pour une classification binaire ressemble à ceci :

| | | Classe prédite | |
|---|---|---|---|
| | | 0 | 1 |
| Classe réelle | 0 | Vrai Négatif | Faux Positif |
| | 1 | Faux Négatif | Vrai Positif |

- Non Défaillant = Classe 0 (Négatif)
- Défaillant = Classe 1 (Positif)

**REMARQUE :**

Vous l'aurez compris, tout l'enjeu du modèle ML est de réussir à minimiser les Faux Positifs mais surtout les Faux Négatifs, c'est-à-dire les personnes qui sont réellement défaillantes et que nous prédisons comme non-défaillantes. En effet, le coût d'une telle erreur pourrait être très élevé pour la banque relativement aux Faux Positifs. Par conséquent, les erreurs de prédiction doivent être minimisées. Dans cette logique, une fonction coût ayant pour objectif de pénaliser les Faux Positifs et les Faux Négatifs sera implémentée et des métriques adéquates seront calculées pour évaluer et comparer la performance des modèles.

[7]Dans une optique de Machine Learning, le prétraitement ou « preprocessing » prépare les données d'entrainement en vue de leur ingestion par le modèle d'apprentissage. Il s'agit d'une étape préliminaire clé qui regroupe un ensemble de phases préparatoires, de la collecte jusqu'aux processus de validation des données. Une des étapes cruciales du processus consiste à explorer les données (dite aussi EDA : Exploratory Data Analysis), les formater, rectifier les éventuelles erreurs qu'elles peuvent comporter, et éventuellement les enrichir (features engineering).

# Application du ML dans la détection de fraude

La détection des fraudes est un défi de taille malgré le fait qu'elle ne représente qu'une très petite fraction de l'activité au sein d'une organisation. En effet, un faible pourcentage de l'activité peut rapidement se transformer en des pertes financières importantes sans parler de la perte de confiance de la clientèle envers l'établissement. A titre d'exemple, en 2020, le montant des fraudes à la carte bancaire en France s'élevait à environ 473 millions d'euros selon la Banque de France. L'augmentation du nombre de transactions frauduleuses s'explique notamment par l'accroissement des canaux de distribution des services financiers via la digitalisation. Par conséquent, le défi pour les acteurs de l'industrie est de mettre en oeuvre une évaluation en temps réel et d'améliorer la précision de la détection des fraudes qu'elles soient externes, internes ou managériales.

Nous allons tenter d'expliquer en quoi le ML est une approche complémentaire voire meilleure dans certains cas comme celui de la détection de paiement par carte. Disons qu'intuitivement, ce qui parait différent devrait attirer notre attention. C'est cette posture qui guide naturellement la lutte contre la fraude du point de vue de la data science. Si nous reprenons l'exemple de la fraude à la carte de crédit, traditionnellement les activités suspicieuses sont détectées en examinant des signaux apparents et évidents. C'est le cas par exemple lorsque, de manière inhabituelle, des transactions importantes se produisent dans des lieux atypiques. En somme, les détections sont programmées selon la règle : Si…Alors. Ainsi, nous pouvons facilement imaginer la limite de ce système dans le traitement des flux de données considérables. La puissance du ML est qu'il est tout d'abord capable de détecter des signaux faibles dans le comportement des utilisateurs et qui ne sont pas évidents. Des signaux synonymes d'une éventuelle fraude. Puisque rappelons-le, le ML traite une grande volumétrie de données, en étudiant les corrélations cachées entre le comportement des clients et la probabilité de transactions frauduleuses. Le secteur bancaire étant par essence au centre des flux de Big Data, l'application du ML permettrait d'améliorer significativement leurs techniques de traitement des données et à terme leurs bénéfices. Ensuite, comme nous venons de le dire : ce qui est différent serait sujet à suspicion. Autrement dit, le premier défi est de définir ce qu'est « la normalité ». Pour ce faire, il faut disposer d'une base historisée de données clients (données bancaires, données de contacts, etc.), la cible pouvant être bien définie (ML supervisé avec une data labellisée) ou non (ML non supervisé ). Ces algorithmes, qu'ils soient supervisés ou non, permettront d'identifier des comportements atypiques et suspects à travers :

► L'automatisation de l'analyse des données et donc d'améliorer la vitesse de détection ;

► La détection des tendances ou patterns des transactions indésirables ;

► L'apprentissage en continu avec une détection en temps réel.

### Si nous sommes dans une approche supervisée :

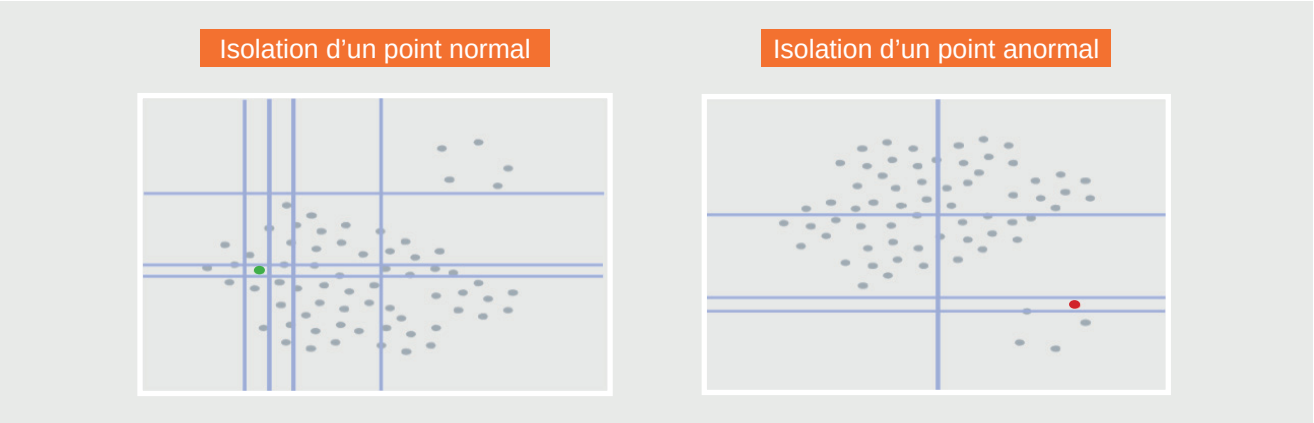
→ Cela veut dire qu'un nombre suffisant de cas de fraudes avérées a été détecté et répertorié (1 : fraude ; 0 : pas de fraude). Nous ne détaillerons pas les phases de nettoyage et préparation des données hormis un point dont l'impact positif sur la performance du système est important : le déséquilibre d'un jeu de données, i.e. le fait d'avoir beaucoup plus de cas non frauduleux que frauduleux, conduit souvent à un mauvais apprentissage par les modèles supervisés. Une fois que sont corrigés[8] les biais résultant de la dissymétrie des données, le ML supervisé présente des performances relativement significatives même si leur explicabilité est souvent sujet à discussion ce qui complique leur application pour des aspects réglementaires. C'est pourquoi dans le paradigme de la détection de fraude, l'approche métier classique et l'approche IA sont souvent complémentaires. L'une apportant du sens à l'autre notamment s'agissant de fournir de nouvelles variables et indicateurs pour l'amélioration continue de ces modèles.

### Si nous sommes dans une approche non supervisée

→ Cela veut dire que les données d'entraînement sont fournies sans étiquettes ni instructions. Dans ce cas, il convient de spécifier ce qu'est la « normalité » ou bien une « anomalie ». Les concepts associés à cette démarche possèdent des définitions qui se confondent plus ou moins suivant les auteurs et les contextes. Nous dirons que nous nous concentrons sur les points aberrants. Il existe plusieurs méthodes pour détecter des anomalies dans un jeu de données. Isolation Forest ou méthode de la forêt d'isolement est l'une des méthodes de détection les plus répandue. Considérons donc le problème de détection de fraude bancaire à partir d'un certain nombre de variables explicatives. L'isolation Forest est une technique qui identifie directement les anomalies (outliers).

[8] Il existe plusieurs solutions comme les méthodes de rééchantillonnages comme l'oversampling ou l'undersampling qui sélectionnent aléatoirement des données afin d'équilibrer le ratio de classe.

C'est une méthode non supervisée. L'idée principale de cet algorithme est qu'une donnée atypique sera plus facile à isoler qu'une donnée standard. Par exemple, si j'ai une base client avec une majorité de clients de moins de 30 ans, alors un client de 60 ans sera facilement isolé grâce à son âge. Le diagramme ci-dessous en illustre le principe :




Le diagramme montre que plus le partitionnement (split) est rapide, plus il y a d'outliers. Autrement dit, les valeurs atypiques sont celles que nous aurons réussi à isoler le plus facilement (point rouge). A contrario, il aura fallu plusieurs splits aléatoires pour isoler des points non atypiques (point vert). Bien entendu, le processus est répété plusieurs fois et le niveau d'isolement pour chaque point/instance est assorti d'un score d'anomalie suggérant sa probabilité d'être une anomalie. Le score est fonction du niveau moyen auquel le point[9] a été isolé. Dans ce domaine, nous pouvons citer l'exemple de Mastercard[10]. En effet, avec plus de 2 milliards de cartes en utilisation dans 330 pays, le groupe peut vérifier jusqu'à 160 millions de transactions par heure, en utilisant des algorithmes de ML et en appliquant plus de 1,9 millions de règles différentes pour examiner chaque transaction. Le tout en quelques millisecondes seulement !

## Pour conclure

A l'intersection des statistiques et de l'informatique, le Machine Learning se préoccupe de la modélisation des données. Dans cet article, nous avons illustré quelques applications du ML dans le secteur bancaire. Une application qui est de plus en plus prévalente et représente un avantage considérable pour les institutions bancaires. Les exemples dans le crédit scoring et la détection de fraude sont parmi les domaines clés dans lesquels les algorithmes de Machine Learning sont d'ores et déjà utilisés pour améliorer les processus de prise de décision, mais pas seulement. Le recours aux algorithmes dans la valorisation des actifs financiers, les notations de dettes d'entreprises ou encore dans la gestion de risques sont autant de disciplines qui constituent un terrain fertile pour le ML. Par conséquent, nous pouvons nous attendre à une adoption accrue de ces technologies nombreuses et variées. Nous avons, à travers des illustrations et exemples d'applications, montré l'intérêt catalyseur des algorithmes à la condition d'en comprendre la mécanique. Choisir un mauvais modèle conduira à de mauvais résultats. Raison pour laquelle, il est crucial de comprendre les étapes préalables afin de pouvoir anticiper et apporter les ajustements et les choix adéquats. Privilégier la compétence à résoudre les problèmes et à assurer une base de données de qualité est primordial si nous voulons aboutir à des prédictions précises. C'est pourquoi dans nos prochaines publications nous aborderons le domaine de la Data Quality et les étapes clés avant la modélisation et la mise en production.■

■ **Mohammed MEHANAOUI**
Responsable d'Études R&D et Innovation | Pôle Transformation, Finance & Digital

[9]Les tops «n» points rassemblés sur la base du score, sont étiquetés comme des anomalies.

[10]« MasterCard : l'intelligence artificielle pour combattre la fraude à la carte bancaire » : mastercard-lintelligence-artificielle-pour-combattre-la-fraude-a-la-carte-bancaire

# RÉFÉRENCES


► Albanesi, S., & Vamossy, D. F. (2019). Predicting consumer default : A deep learning approach (No. w26165). National Bureau of Economic Research.

► Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of fintechs : Credit scoring using digital footprints. The Review of Financial Studies, 33(7), 2845-2897.

► Chesneau, D. Fintechs et banques : coopération et coopétition (2019). Annales des Mines - Réalités industrielles. pp. 45-50.

► Denis, N. La banque, en pleine transformation (2019). Annales des Mines - Réalités industrielles. (2019). pp. 33-35.

► Hurlin, C. & Pérignon, C. (2019). Machine learning et nouvelles sources de données pour le scoring de crédit. Revue d'économie financière, 135, 21-50.