nexialog CONSULTING

BENCHMARK

Des solutions de data anonymisation

Yann HUQUET

TABLE DES MATIÈRES

INTRODUCTION

Les données synthétiques, comme leur nom l'indique, sont des données créées artificiellement par des programmes d'Intelligences Artificielles (IA). Il peut s'agir de textes, d'images, de voix ou de séquences vidéo. Ces données synthétiques peuvent être appliquées à presque tous les secteurs : informatique et logiciels, commerce de détail, finance, défense, soins de santé, agriculture, production alimentaire, construction, jeux, et bien d'autres encore.

L'utilisation des techniques d'anonymisation permet de résoudre un problème majeur : **la protection des données personnelles**.

Bien que certaines banques fournissent à leur équipe des API visant à partager leur données tout en conservant la confidentialité et la transparence des données, travailler sur les données synthétiques permet de sécuriser davantage les processus et de respecter le RGPD, ensemble de règles protégeant les données des utilisateurs et leur transparence. Le règlement général sur la protection des données est disponible sur le site de la CNIL [1].

En effet, lorsqu'un pirate s'introduit dans un système, il peut récupérer des données sensibles de centaines de clients. L'accès et l'utilisation de ces données sensibles sont donc limités aux entreprises et aux personnes concernées.

Pour surmonter ce problème, les entreprises se tournent donc vers des outils de génération de données anonymisées. Ces données offrent un moyen alternatif de capturer des informations du monde réel, et peuvent être utilisées ou partagées à moindre risque. Cependant, cet aspect doit bien être nuancé. En effet si la base synthétisée présente les mêmes propriétés statistiques que la base réelle, les analyses qui en découlent peuvent également véhiculer des informations sensibles.

Cette étude a pour but de faire une revue des approches possibles du sujet et de comparer les solutions disponibles pour anonymiser les données structurées.

L'anonymisation des données sera définie, puis les acteurs du marché seront comparées selon des critères préalablement établis.

I. Pourquoi et comment anonymiser des données ?

1.1. Pourquoi anonymiser des données ?

Le partage des données fait partie des exigences opérationnelles des institutions bancaires. En effet, la mise en commun d'informations rend possible de meilleures analyses clients grâce à la centralisation des savoirs. Ce partage permet ainsi de comprendre les comportements des clients en fonction d'informations tirées de leurs transactions par exemple.

La vraie question est : pourquoi ne pas simplement utiliser des données réelles ? Une raison est le manque de contrôle sur les données ainsi que leur protection.

En effet, le partage des données à l'intérieur de l'entreprise soulève certaines problématiques. Même si les données personnelles sont nécessaires aux conseillers opérationnels, utiliser des données personnelles pour de la gestion de risques n'est pas nécessaire et peut induire un risque de biais dans les modèles.

Il s'agit autant d'un enjeu de risque de modèles que d'un enjeu de confidentialité des données.

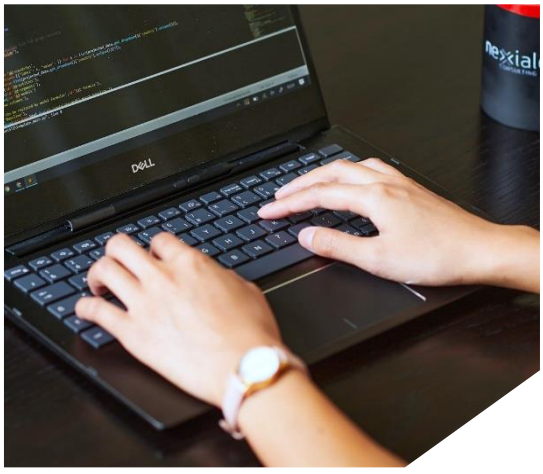
De plus, dans le cadre d'un projet, le transfert des données vers une personne externe au groupe ou même vers une autre entité du groupe pose problème d'un point de vue du RGPD.

La différence entre Data Anonymisation, Data Pseudonymisation et Data Synthétisation

L'anonymisation des données vise à protéger les informations privées ou sensibles d'un individu en effaçant ou en cryptant les identifiants qui le relient aux données stockées, tout en conservant une information pertinente et peu dégradée pour le cas d'usage visé.

Cependant, même lorsque ces identifiants sont supprimés ou déplacés, des méthodes de désanonymisation, pour retracer le processus d'anonymisation des données, peuvent croiser plusieurs sources d'information et révéler des informations personnelles.

Plusieurs techniques d'anonymisation des données existent, dont la pseudonymisation et la synthétisation de données.



La pseudonymisation des données

La pseudonymisation est une méthode de gestion et de dépersonnalisation des données qui remplace les identifiants privés par de faux identifiants (ou “pseudonymes”), par exemple en remplaçant l'identifiant "John Smith" par “Mark Spencer”.

La pseudonymisation préserve la précision statistique et l'intégrité des données, ce qui permet d'utiliser les données modifiées pour la formation, le développement, les tests et l'analyse.

Cependant, les données dépersonnalisées peuvent être réassociées aisément à l’identifiant dont elles proviennent. En effet, les informations nécessaires à cette fin (les pseudonymes synthétiques) peuvent être réassociées aux données sources et doivent donc être conservées séparément et en sécurité pour éviter toute violation de la vie privée.

Davantage de détails sur les différences de définitions entre ces termes sont disponibles dans l’article de J. Arthur [2].

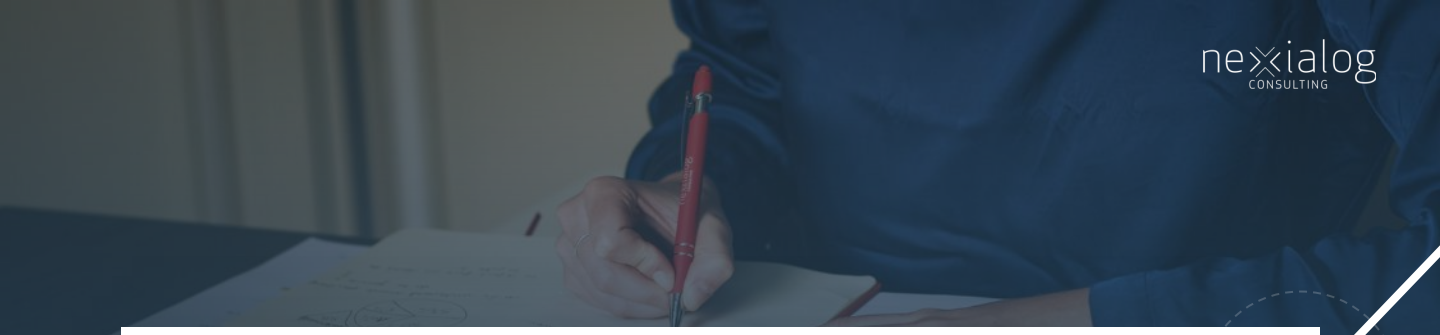
La synthétisation des données

La génération de données synthétiques est un processus mathématique et statistique réalisé par des modèles d'apprentissage automatique entraînés à partir d'un environnement réel.

Les données de sortie ne contiennent pas de données sensibles mais conservent néanmoins les caractéristiques comportementales des données réelles (conservation des distributions des variables, des corrélations entre variables, des résultats à certains modèles, etc.).

La génération des données synthétiques n’est pas aisée puisqu’il ne s’agit pas simplement de remplacer les données sensibles (numéro de comptes, contrats, etc.) mais bien de fabriquer un nouveau jeu de données ayant les mêmes propriétés statistiques que les données réelles.

Contrairement à la pseudonymisation des données, **la synthétisation de données permet de conserver l’information disponible dans la table initiale** ainsi que sa forme (même nombre de variables, mêmes corrélations entre les variables, etc.).



Les avantages de la donnée synthétique

La sécurité des données synthétisées est un de leurs avantages majeurs. En effet, une fois les données clients synthétisées en main, il est tout simplement impossible de remonter aux données originelles (réelles).

La génération de données synthétiques par apprentissage automatique permet également d'en générer la quantité souhaitée. Il est alors possible de compléter des ensembles de données qui ne comporteraient pas suffisamment d'exemples.

En règle générale, des données d'entraînement plus nombreuses et de meilleure qualité sont synonymes de meilleures performances pour les modèles. Les données synthétiques peuvent donc jouer un rôle extrêmement important pour les ingénieurs travaillant dans des domaines où les données sont rares. Le terme de « data augmentation » est alors employé.

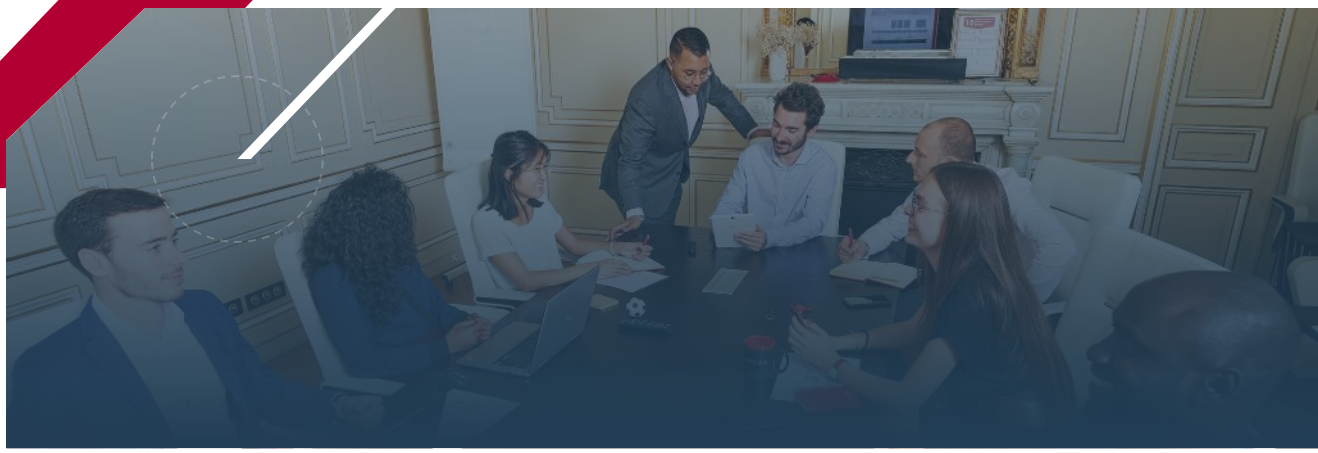
Data augmentation et data reduction

La data augmentation consiste donc à augmenter la quantité de données disponibles pour un apprentissage. En effet, l'entraînement d'un modèle sur peu d'observations peut engendrer un phénomène de sur-apprentissage.

Le sur-apprentissage apparaît lorsque le modèle est entraîné sur l'échantillon d'entraînement, sans pour autant être capable d'émettre des prédictions pertinentes sur de nouvelles données. Dans ce cas, les performances de modèles sont faibles sur l'échantillon de test alors qu'elles étaient bonnes sur l'échantillon d'entraînement.

Ce phénomène est résolu le plus souvent en augmentant la taille de l'échantillon d'apprentissage et/ou en réduisant le nombre de paramètres du modèle.

En augmentant la taille du jeu de données, tout en réduisant le nombre d'erreurs et en équilibrant les données (pour avoir une classe minoritaire davantage représentée), les outils génèrent ainsi des données d'entraînement synthétiques optimales avec un biais plus faible.



Les bénéfices de la data augmentation dans les réseaux de neurones sont développés davantage dans l'article de Hernández-García, A., & König, P. (2018) [3].

La réduction de la taille du dataset pourrait également être envisagée car elle permet, pour d'importants jeux de données, d'accroître l'efficacité du stockage et de réduire les coûts (de gestion de base ou de calcul par exemple).

1.2. Comment synthétiser les données ?

Afin de synthétiser les données pour les raisons évoquées précédemment, telles que le transfert interne ou externe des données sensibles ou leur utilisation pour des modèles, plusieurs algorithmes existent. Basées sur le principe des réseaux de neurones, les méthodes disponibles et adaptées au sujet sont multiples. Ce point est la principale différence entre les outils du marché.

Les réseaux de neurones Variational Auto-Encoder (VAE) et Generative Adversarial Network (GAN) sont deux réseaux de neurones utilisés pour la synthétisation des données. Leurs caractéristiques respectives et leurs différences sont détaillées ci-après.

Qu'est ce qui différencie VAE et GAN ?

La principale différence entre les VAE et les GAN est leur processus d'apprentissage.

Les **VAE** utilisent une architecture de réseaux de neurones constituée d'un encodeur, qui compresse la donnée d'entrée en une distribution de probabilité, et d'un décodeur, qui reconstruit la donnée encodée. Ils peuvent donc être considérés comme résolvant un problème d'apprentissage semi-supervisé. L'algorithme étudie des distributions et autres caractéristiques puis mémorise ces points clés pour reconstruire un jeu de données synthétique.

Le processus du modèle VAE est schématisé p.8.

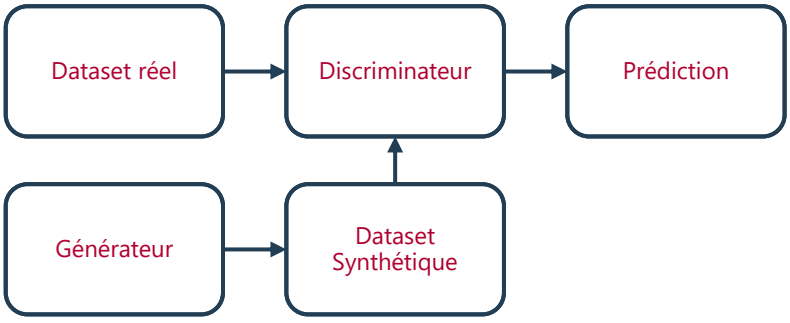
Kingma, D. P., & Welling, M. (2014) [6] s'intéressent, en détail à l'utilisation des VAE.

VAE

Les **GAN** entraînent deux réseaux neuronaux, respectivement appelés générateur et discriminateur. Le générateur crée des données synthétiques (apprentissage non supervisé), et le discriminateur essaie de déterminer si le résultat est réel ou non (apprentissage supervisé). Ce processus fonctionne en boucle, et la qualité des données produites par le générateur ne cesse de s'améliorer car le générateur apprend grâce au discriminateur et réduit les différences entre les jeux de données réels et synthétiques.

Finalement, le discriminateur devient incapable de faire la différence entre les données réelles et les données synthétiques. La discrimination est bien supervisée car la comparaison est effectuée par rapport au jeu de données réel, labellisé.

Le temps d'apprentissage des GAN est plus long que celui des VAE, car comprenant une partie génération puis une partie discrimination. En utilisant les VAE, il est possible d'obtenir des résultats plus rapides, mais parfois moins performants, par rapport aux résultats des GAN, l'encoder étant pré-entraîné.



GAN

Vaseekaran, V. (2022) [4] ou Vieira, A. (2023) [5] donnent davantage de détails sur les réseaux de neurones GAN et leur utilisation lors de la synthétisation de données.

II. Comparaison des différentes solutions du marché

Une multitude d'entreprises ont fait de la synthétisation de données leur spécialité. Cependant, le sujet étant vaste et les possibilités de développement et d'axes de travail étant multiples, ces solutions se différencient selon plusieurs critères, énoncés ci-après.

Le domaine d'expertise et type de données

Dans un premier temps, bien que plusieurs entreprises soient spécialisées dans la synthétisation de données, les secteurs d'activités visés et les types de données traitées diffèrent.

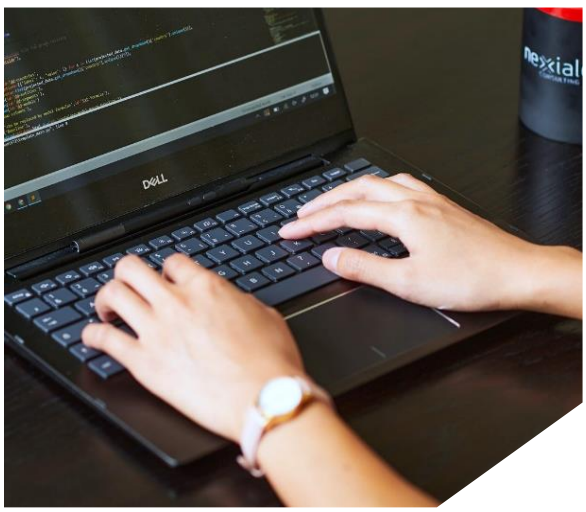
Cette étude se concentre sur les solutions d'anonymisation de données structurées et ciblant les secteurs bancaires et assurantiels.

Hazy, **Datomize**, **Mostly AI**, **Ydata** et **Tonic**, notamment, sont spécialisées dans la génération de données synthétiques. Ces solutions visent à anonymiser des données afin de partager l'information pertinente en filtrant l'information sensible mais également pour construire ou tester des modèles bancaires ou assurantiels.

Ces entreprises participent à la mise en conformité de leur client par rapport aux réglementations de protection des données (RGPD, E-Privacy Regulation, Data Act, etc.).

D'autres entreprises synthétisent des données mais ne traitent que des images, sons ou textes dans les secteurs de l'imagerie satellitaire ou de la télédétection de défense. Ces outils ne seront donc pas étudiées ici.

Ainsi, une première caractérisation des solutions du marché est possible simplement par leur domaine d'expertise principal et le type de données traitées.



Les étapes d'un projet de synthétisation

D'un point de vue technique, les étapes utilisées pour générer les données synthétiques sont en général les mêmes pour tous les outils du marché, cependant ce sont les méthodes employées et les livrables rendus qui diffèrent. En effet, chaque outil reçoit le jeu de données à synthétiser, le synthétise, puis le compare par rapport au jeu de données réel.

Dans la plupart des cas, les outils développés présentent une partie de préparation et de paramétrisation de la synthétisation du jeu de données, couplée à une visualisation d'indicateurs de performances.

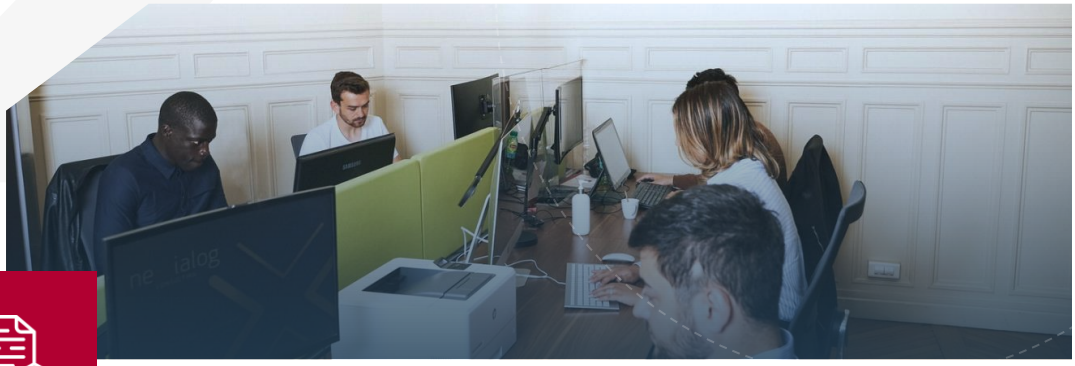
Certaines entreprises, telles que **Hazy** ou **Datomize** par exemple, proposent une analyse préalable du dataset à synthétiser.

La génération du dataset synthétique

La génération du dataset synthétique peut être réalisée de différentes manières, selon les solutions étudiées et selon les données fournies. En effet, l'approche est différente selon les données à traiter. Les algorithmes de **Datomize** ou **Tonic** sont pré-entraînés puis appliqués aux données d'entrée ; d'autres algorithmes sont spécifiques à chaque cas d'usage, c'est le cas de **Hazy** ou **Mostly AI**, où les modèles apprennent des données d'entrée et de leur caractéristiques afin de développer une solution adaptée à chaque jeu de données.

Les méthodes de synthétisation des entreprises s'appuient sur deux familles de réseaux de neurones, présentées en Section 1.2 : les GAN (Generative Adversarial Network) et les VAE (Variational Auto-Encoder).

- **Hazy** : GAN
- **Datomize** : VAE
- **Mostly AI** : GAN
- **YData** : GAN
- **Tonic** : VAE



Comparaison des jeux de données réels et synthétiques

Comment savoir si les données synthétiques conservent la même richesse d'information, les mêmes corrélations et les mêmes propriétés que les données originales ?

Comment être sûr que les données synthétiques sont vraiment anonymisées et ne peuvent pas faire l'objet d'une rétro-ingénierie pour divulguer des informations privées ?

Ces questions sont primordiales lorsque la synthétisation de données est mise en place. Chaque outil utilise donc une multitude de métriques lui permettant de juger notamment de la similarité et de la confidentialité de la synthétisation réalisée.

Tous les outils sans exception comparent les propriétés statistiques du jeu d'entrée et de sortie. Les distributions marginales des variables (à l'aide d'histogrammes ou box-plots), les corrélations entre les variables et la fidélité de couverture (même nombre de valeurs manquantes, de valeurs aberrantes) sont étudiées.

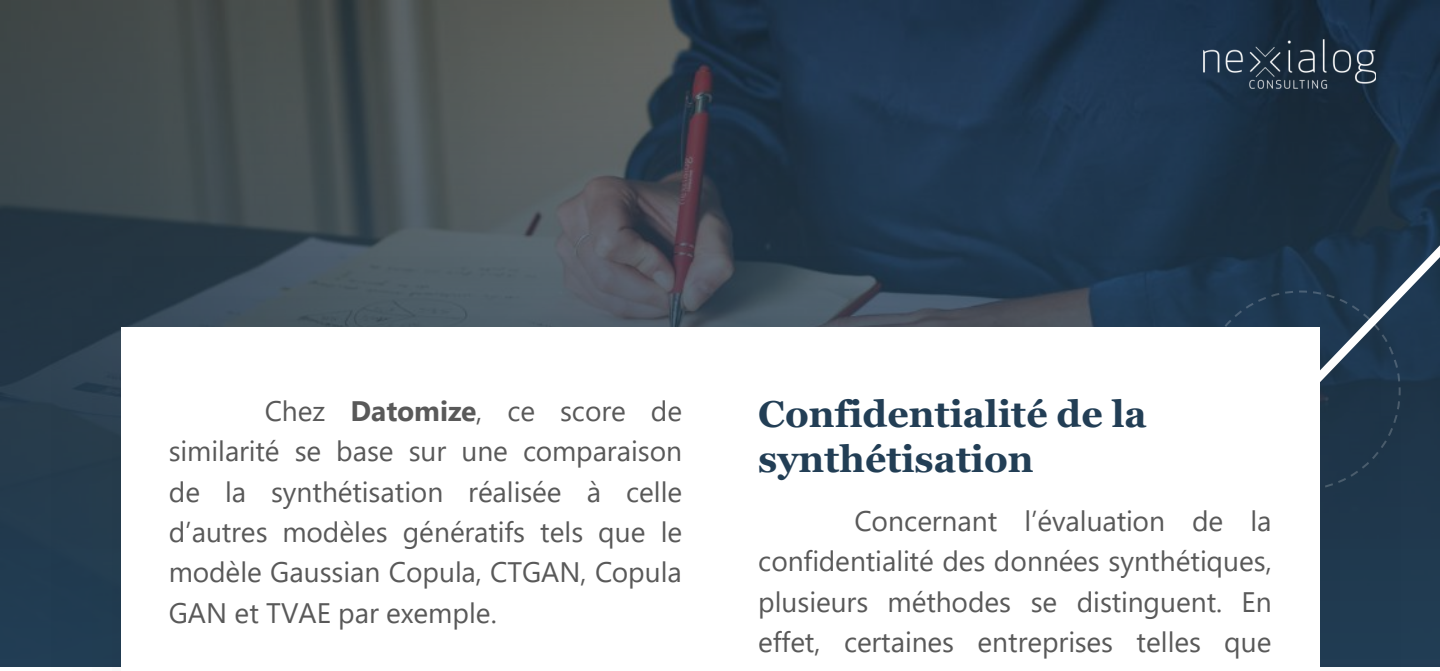
Les modèles d'apprentissage automatique des différentes solutions du marché sont évalués en fonction de trois critères de référence clés : **fidélité**, **confidentialité** et **utilité**.

La fidélité mesure la correspondance statistique entre les données synthétiques et les données d'origine, la confidentialité indique le niveau d'anonymisation des données synthétiques et l'utilité indique la performance des données synthétiques dans leurs applications par rapport à l'ensemble de données d'origine.

Ces deux objectifs de similarité statistique et de confidentialité sont antagonistes car l'augmentation de l'un a tendance à diminuer l'autre. En pratique, les exigences du cas d'utilisation des données synthétiques déterminent la valeur à accorder à la similarité statistique par rapport à la confidentialité.

Fidélité de la synthétisation

Un score de similarité est notamment créé par **Datomize** et **YData** et permet d'obtenir un aperçu rapide de la performance globale de l'outil en matière de fidélité.



Chez **Datomize**, ce score de similarité se base sur une comparaison de la synthétisation réalisée à celle d'autres modèles génératifs tels que le modèle Gaussian Copula, CTGAN, Copula GAN et TVAE par exemple.

Le score évaluateur de table est basé sur les mesures suivantes :

- la similitude des indicateurs statistiques de base (moyenne, médiane, écart type et variance) des échantillons source et synthétique ;
- la corrélation colonne par colonne entre les deux versions de données (corrélation inter-tables), basée sur la corrélation de Pearson pour les colonnes numériques et le V de Cramer pour les colonnes catégorielles ;
- des tests statistiques sont exécutés sur toutes les colonnes compatibles (ainsi, les distributions des colonnes catégorielles ou booléennes sont comparées avec le test du Khi-deux et les distributions des colonnes numériques sont comparées avec le test de Kolmogorov-Smirnov à deux échantillons).

La méthodologie de scoring de **Datomize** est explicité davantage par Fried, G. (2022) [7].

Confidentialité de la synthétisation

Concernant l'évaluation de la confidentialité des données synthétiques, plusieurs méthodes se distinguent. En effet, certaines entreprises telles que **Hazy** par exemple, s'appuient sur la confidentialité différentielle tandis que d'autres telles que **Datomize** ou **Tonic** tentent de détecter les données synthétiques des données réelles. En mélangeant les données réelles et les données synthétiques avec des flags indiquant si ces données sont réelles ou synthétiques, des classificateurs sont construits et tentent de prédire ce flag. Le but étant, pour un individu tiré aléatoirement du jeu réel ou synthétique, de ne pas être en mesure de déterminer le jeu de données dont il est issu.

La confidentialité différentielle est une garantie forte, mathématiquement prouvable, de la protection de la confidentialité. Elle permet d'extraire des informations utiles d'ensembles de données contenant des informations personnelles et d'offrir une meilleure protection de la vie privée. Elle est obtenue en introduisant un "bruit statistique". Ce bruit est suffisamment important pour protéger la vie privée de tout individu, mais suffisamment faible pour ne pas affecter la précision des réponses extraites par les analystes et les chercheurs.

Dwork, C., & Roth, A. ont publié un article sur les fondements de la confidentialité différentielle en 2014 [8].

La solution proposée par **Hazy** garantit la confidentialité différentielle en rendant son générateur de données agnostique à la présence (ou non) de tout enregistrement individuel dans les données sources. Elle y parvient en ajoutant du bruit à la forme des distributions dans son générateur. Cela rend ses générateurs différentiellement privés par conception et permet de gérer aisément le niveau de confidentialité requis.

L'outil met également en œuvre une confidentialité différentielle qui garantit qu'aucune donnée originale n'est mémorisée ou réidentifiée dans le système.

Datomize évaluent la difficulté de distinguer les données synthétiques des données réelles en utilisant un modèle d'apprentissage automatique (Logistic Detection ou SVC). Pour ce faire, les données sont marquées comme réelles ou synthétiques, puis regroupées ; un modèle tente ensuite de les classer comme réelles ou synthétiques. Le but est alors que le modèle ne soit pas en mesure de différencier l'origine des enregistrements, ici, le F1-score du modèle permet d'évaluer la confidentialité de la nouvelle base de données.

Chez **YData** ou **Mostly AI**, les scores de correspondance, de confidentialité des voisins et d'inférence jugent la confidentialité des données synthétiques.

Ce score de correspondance compte le nombre d'enregistrements sensibles dans les données synthétiques qui correspondent aux enregistrements de l'ensemble de données original. Le score de confidentialité des voisins mesure la probabilité de trouver les points de données originaux dans un certain rayon autour des points de données synthétiques. Ce score est calculé en effectuant une recherche des plus proches voisins à haute dimension sur les données synthétiques superposées aux données originales.

Le score d'inférence mesure la probabilité qu'un attaquant puisse déterminer l'origine d'une données en particulier. Le but est de ne pas être en mesure de différencier les données réelles des données synthétiques.

Utilité de la synthétisation

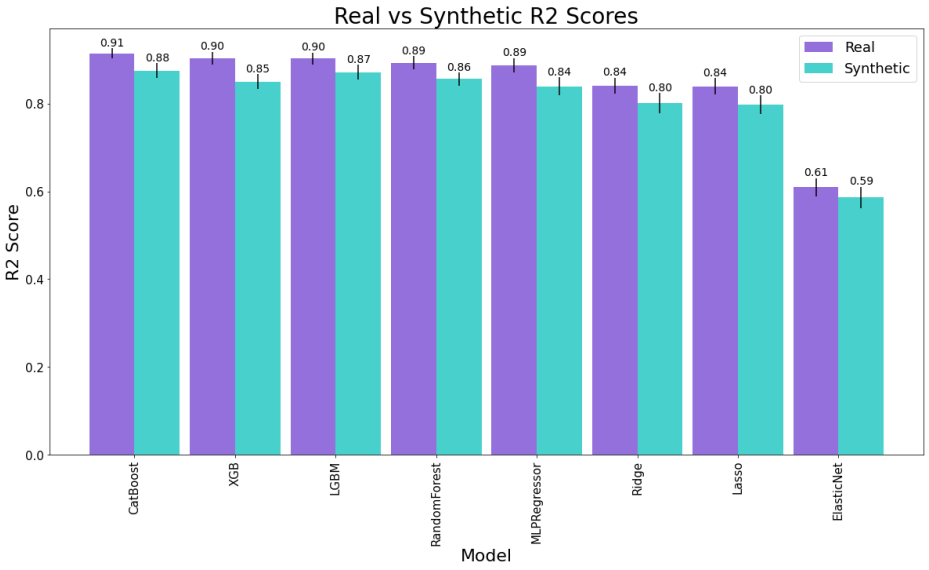
L'utilité s'appuie sur la comparaison des résultats des modèles de Machine Learning basés sur des données réelles et synthétiques. L'évaluation est faite sur des données réelles alors que les modèles comparés sont entraînés sur des données réelles et synthétiques.

Chez **YData** par exemple, un Model Score est créé pour les données réelles et synthétiques. Le score comprend plusieurs modèles pour les tâches de régression ou de classification. Si les scores sont comparables, cela indique que les données synthétiques ont la qualité nécessaire pour être utilisées pour entraîner des modèles performants pour des applications du monde réel.

Un score d'importance des variables (Features Importances) prolonge le score du modèle. De l'interprétabilité des modèles formés est ajouté en mesurant les changements dans l'importance de chaque variable pour un modèle formé sur des données synthétiques et un modèle formé sur l'ensemble de données original.

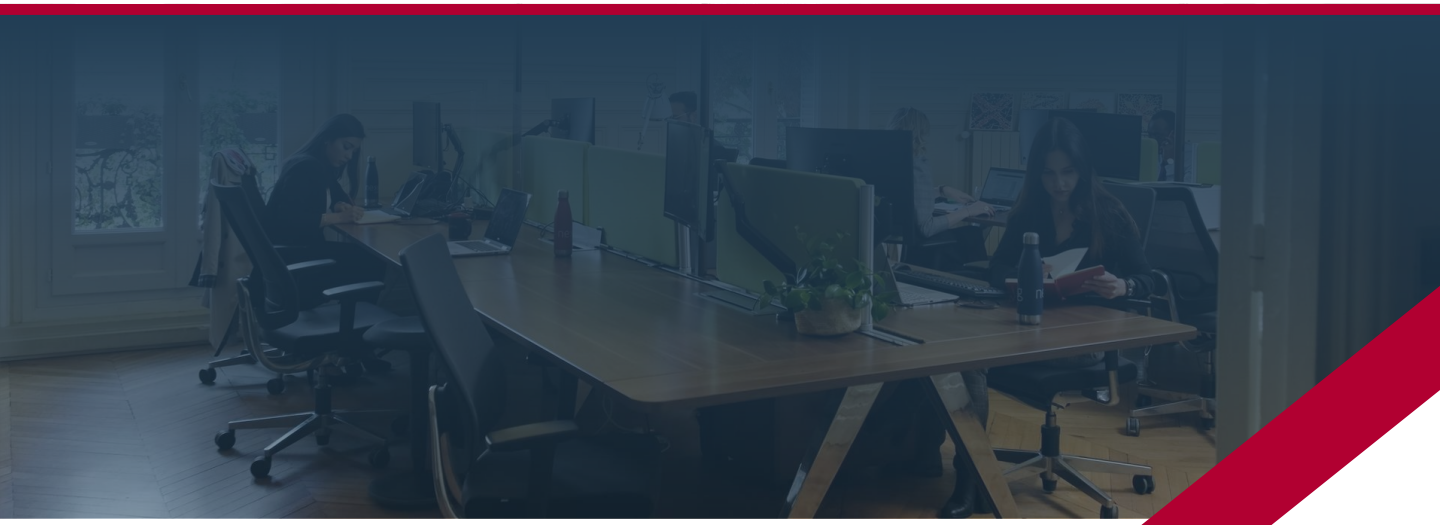
Comparer les performances d'un modèle sur les données synthétiques et celles de ce même modèle sur les données réelles, permet de déceler les différences de comportements des deux jeux de données, le but de la synthétisation de données étant d'obtenir un jeu de données ayant le même comportement face aux modèles.

Mostly AI ou **Tonic** utilisent également les données synthétiques afin d'améliorer l'accuracy de certains modèles de Machine Learning [9].



Comparaison des performances des échantillons réels et synthétiques pour des modèles de Machine Learning selon le R^2. ***Tonic*** *[10]*

Dans les solutions étudiées, d'autres points de comparaison entre données réelles et synthétiques existent, telles que la présence de données sources dans les données synthétiques.



Accès aux solutions

Finalement les accès aux solutions de synthétisation varient : qu'elles soient en API (comme **Datomize**, **Tonic**, **Mostly AI** ou **Gretel**), sur un cloud (comme **Hazy**), ou sur site (comme **Hazy** ou **Mostly AI**).

Le prix et l'assistance fournis, qui sont cruciaux, permettent la différenciation de ces solutions. Aucun chiffre ne sera donné ici, simplement car les entreprises adaptent leur politique tarifaire au cas d'usage présenté ainsi qu'à l'utilisation qu'en fait le client ou l'assistance fournie par exemple.

Tableau récapitulatif des comparaisons

	Hazy	Datomize	Mostly AI	YData	Djinn Tonic
Secteur d'activité bancaire et assurantiel	☑	☑	☑	☑	☑
Etude et exploration préalable du dataset input	☑	☑	Aucune visualisation ou analyse du dataset réel		
Accès à la solution	Cloud ou sur site	API	API, on site	Cloud	API
Synthétisation des données					
Méthodes de synthétisation	GAN	VAE	GAN	GAN	VAE
Apprentissage du modèle de synthétisation	Les modèles apprennent des données d'entrée	Modèles pré-entrainés	Les modèles apprennent des données d'entrée	Modèles pré-entrainés	Modèles pré-entrainés
Data augmentation / reduction	☑	☑	☑	☑	☑
Comparaison des jeux de données réels et synthétiques					
Critères de fidélité scorés	Comparaison visuelle	☑	Comparaison visuelle et scorée	☑	Comparaison visuelle et scorée
Comparaison de ML	☑	☑	☑	☑	☑
Critères de confidentialité	Confidentialité différentielle	Détection des données réelles / synthétiques	Score de confidentialité par l'étude des proches voisins	Score de confidentialité par l'étude des proches voisins	Détection des données réelles / synthétiques

Conclusion

Pour conclure, cette revue revient sur la définition et les enjeux de l'anonymisation des données ainsi que les différentes solutions disponibles sur le marché.

Les secteurs bancaires et assurantiels font face au durcissement des réglementations de protections des données sensibles, telles que les numéros de comptes, de contrats, etc. Anonymiser les données par la synthétisation est une solution visant à sécuriser la donnée tout en conservant la totalité des informations contenues dans les bases et semble être la solution la plus adaptée au problème posé.

Ce sujet d'anonymisation soulève, en effet, une problématique de maximisation de la confidentialité et de l'utilité des données sous contrainte de la fidélité. Fidélité et confidentialité étant deux objectifs discordants, la volonté de chaque entreprise est de trouver le meilleur compromis entre ces deux caractéristiques.

Utiliser les données synthétiques permet d'éliminer les risques liés à la confidentialité en réduisant au minimum l'interaction avec les données réelles des clients, grâce à l'anonymisation irréversible de l'IA, tout en conservant parfaitement leur structure, corrélations et dépendances temporelles et ainsi être plus précis.

Enfin, il est possible d'améliorer les performances des modèles de Machine Learning en augmentant l'échantillonnage des événements rares en générant des données synthétiques plus équilibrées (par exemple, avec plus de cas de fraude). Ceci augmente la précision des modèles en aval et limite le sur-apprentissage.

Plusieurs entreprises gèrent la synthétisation de jeux de données ainsi qu'une comparaison avec le jeu de données initial. Chaque entreprise se distingue par une méthodologie ou une vision du problème différente des autres.

Annexes & références

Références

1. Le règlement général sur la protection des données - RGPD | CNIL. (s. d.). https://www.cnil.fr/fr/reglement-europeen-protection-donnees
2. Arthur, J. (2023). Anonymisation vs Pseudonymisation vs Synthetic Data. Hazy. https://hazy.com/resources/2019/03/01/anonymisation-vs-pseudonymisation-vs-synthetic-data
3. Hernández-García, A., & König, P. (2018). Further Advantages of Data Augmentation on Convolutional Neural Networks. Artificial Neural Networks and Machine Learning – ICANN 2018, 95-103. https://doi.org/10.1007/978-3-030-01418-6_10
4. Vaseekaran, V. (2022). GANs for Synthetic Data Generation. https://ydata.ai/resources/gans-for-synthetic-data-generation
5. Vieira, A. (2023). The Beauty of GANs : Sharing insights without sharing data. Hazy. https://hazy.com/resources/2020/08/11/the-beauty-of-gans-sharing-insights-without-sharing-data
6. Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. International Conference on Learning Representations. http://pure.uva.nl/ws/files/2511146/162970_1312.6114v10.pd.pdf
7. Fried, G. (2022). Generative Models Benchmark. Datomize. https://www.datomize.com/generative-models-benchmark/
8. Dwork, C., & Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. Foundations and Trends® in Theoretical Computer Science, 9(3-4), 211-407. https://doi.org/10.1561/0400000042
9. Platzer, M. (2022). Boost your Machine Learning Accuracy with Synthetic Data. MOSTLY AI. https://mostly.ai/blog/boost-machine-learning-accuracy-with-synthetic-data/
10. Ferrara, J. (2021). Using Neural Networks to Synthesize Complex Data Relationships, with AI Synthesizer. Tonic https://www.tonic.ai/blog/using-neural-networks-to-synthesize-complex-data-relationships-with-ai-synthesizer

Nexialog Consulting

STRATÉGIE | ACTUARIAT | GESTION DES RISQUES | DATA

Nexialog Consulting est un cabinet de conseil spécialisé en Stratégie, Actuariat, Gestion des risques et Data qui dessert aujourd'hui les plus grands acteurs de la banque et de l'assurance. Nous aidons nos clients à améliorer de manière significative et durable leurs performances et à atteindre leurs objectifs les plus importants.

Les besoins de nos clients et les réglementations européennes et mondiales étant en perpétuelle évolution, nous recherchons continuellement de nouvelles et meilleures façons de les servir. Pour ce faire, nous recrutons nos consultants dans les meilleures écoles d'ingénieur et de commerce et nous investissons des ressources de notre entreprise chaque année dans la recherche, l'apprentissage et le renforcement des compétences.

Quel que soit le défi à relever, nous nous attachons à fournir des résultats pratiques et durables et à donner à nos clients les moyens de se développer.

CONTACTS

Ali BEHBAHANI
Associé, Fondateur
+ 33 (0) 1 44 73 86 78
abehbahani@nexialog.com

Christelle BONDOUX
Associée, Direction Commerciale & Recrutement
+ 33 (0) 1 44 73 75 67
cbondoux@nexialog.com

Vivien BRUNEL
Associé, Data & Innovation
+ 33 (0) 6 71 23 38 97
vbrunel@nexialog.com

Paul-Antoine DELETOILLE
Sales Leader
+33 (0)1 44 73 75 70
+33 (0)7 64 57 86 69
padeletoille@nexialog.com

Areski COUSIN
Directeur Scientifique
acousin@nexialog.com

www.nexialog.com

Retrouvez toutes nos publications sur Nexialog R&D