Haute École

« ICHEC – ECAM – ISFSC »

Enseignement supérieur de type long de niveau universitaire

Comment l'analyse des données peut-elle être utilisée pour cibler efficacement les clients, et personnaliser les offres dans les prochaines campagnes marketing des banques ?

Mémoire présenté par :

Hind BELCAID

Pour l'obtention du diplôme de :

Master en gestion de l'entreprise

Année académique 2023-2024

Promotrice :

Martine GEORGE

Boulevard Brand Whitlock 6 – 1150 Bruxelles

Remerciements

Avant toute chose, j'aimerais remercier un certain nombre de personnes qui m'ont aidé directement et indirectement pour la rédaction de mon mémoire de fin d'étude.

J'aimerais commencer par remercier Mme Martine George qui fut ma promotrice pour ce mémoire. Je tiens à exprimer ma gratitude pour son soutien et ses conseils avisés tout au long de la réalisation de ce travail. Son expertise a été un élément essentiel dans l'accomplissement de ce projet.

Dans un deuxième temps, j'aimerais remercier M. Van Droogenbroeck Vincent pour son encadrement et ses conseils durant mon stage chez Accenture. Son approche bienveillante et ses feedbacks m'ont permis de progresser significativement dans mes compétences professionnelles.

Ensuite, je remercie mes professeurs pour les connaissances qui m'ont été transmises. Leur enseignement rigoureux et passionné m'a permis d'acquérir les compétences nécessaires pour rédiger ce mémoire.

Pour terminer, je souhaite également remercier mes amis et ma famille qui m'ont soutenu non seulement durant la rédaction du mémoire mais durant toutes mes années scolaires à l'ICHEC. Leurs encouragements ont été une source de motivation inestimable.

Engagement anti-plagiat

« Je soussignée, BELCAID Hind étudiante en dernière année de Master 2 en gestion de l'entreprise, déclare par la présente que le mémoire ci-joint est exempt de tout plagiat et respecte en tous points le règlement des études en matière d'emprunts, de citations et d'exploitation de sources diverses signé lors de mon inscription à l'ICHEC, ainsi que les instructions et consignes concernant le référencement dans le texte respectant la norme APA, la bibliographie respectant la norme APA, etc. mises à ma disposition sur Moodle.

Sur l'honneur, je certifie avoir pris connaissance des documents précités et je confirme que le Mémoire présenté est original et exempt de tout emprunt à un tiers non-cité correctement. »

Dans le cadre de ce dépôt en ligne, la signature consiste en l'introduction du mémoire via la plateforme ICHEC-Student.

Déclaration sur l'honneur sur le respect des règles de référencement et sur l'usage des IA génératives dans le cadre du mémoire

Je soussignée, BELCAID Hind, étudiante en dernière année de Master 2 en gestion de l'entreprise, déclare par la présente que le travail ci-joint respecte les règles de référencement des sources reprises dans le règlement des études en signé lors de mon inscription à l'ICHEC (respect de la norme APA concernant le référencement dans le texte, la bibliographie, etc.) ; que ce travail est l'aboutissement d'une démarche entièrement personnelle; qu'il ne contient pas de contenus produits par une intelligence artificielle sans y faire explicitement référence. Par ma signature, je certifie sur l'honneur avoir pris connaissance des documents précités et que le travail présenté est original et exempt de tout emprunt à un tiers non-cité correctement.»

Date : 19/08/2024
Signature : BELCAID Hind

Je soussignée BELCAID Hind 190463, déclare sur l'honneur les éléments suivants concernant l'utilisation des intelligences artificielles (IA) dans mon travail mémoire :

Type d'assistance		**Case à cocher**
Aucune assistance	J'ai rédigé l'intégralité de mon travail sans avoir eu recours à un outil d'IA générative.	
Assistance avant la rédaction	J'ai utilisé l'IA comme un outil (ou moteur) de recherche afin d'explorer une thématique et de repérer des sources et contenus pertinents.	
Assistance à l'élaboration d'un texte	J'ai créé un contenu que j'ai ensuite soumis à une IA, qui m'a aidé à formuler et à développer mon texte en me fournissant des suggestions.	
	J'ai généré du contenu à l'aide d'une IA, que j'ai ensuite retravaillé et intégré à mon travail.	
	Certains parties ou passages de mon travail/mémoire ont été entièrement été générés par une IA, sans contribution originale de ma part.	
Assistance pour la révision du texte	J'ai utilisé un outil d'IA générative pour corriger l'orthographe, la grammaire et la syntaxe de mon texte.	
	J'ai utilisé l'IA pour reformuler ou réécrire des parties de mon texte.	
Assistance à la traduction	J'ai utilisé l'IA à des fins de traduction pour un texte que je n'ai pas inclus dans mon travail.	
	J'ai également sollicité l'IA pour traduire un texte que j'ai intégré dans mon mémoire.	
Assistance à la réalisation de visuels	J'ai utilisé une IA afin d'élaborer des visuel, graphiques ou images.	
Autres usages		

Je m'engage à respecter ces déclarations et à fournir toute information supplémentaire requise concernant l'utilisation des IA dans mon tmémoire, à savoir :
J'ai mis en annexe les questions posées à l'IA et je suis en mesure de restituer les questions posées et les réponses obtenues de l'IA. Je peux également expliquer quel le type d'assistance j'ai utilisé et dans quel but.

Fait à Bruxelles, le 19/08/2024 Signature: BELCAID Hind 190463

“In God we trust, all others bring data”
(W. Edwards Deming, s.d)

TABLE DES MATIÈRES

Liste des Figures :

Liste des Tableaux :

****Remarque :*** *Toutes les figures et tous les tableaux présent dans le chapitre 3 « ANALYSE EXPLORATOIRE ET MODÉLISATION PRÉDICTIVE DES DONNÉES » ont été créées par l'auteur à partir des modèles prédictifs et des données analysées.*

1 Introduction

1.1 Contexte : L'Impact du numérique dans le secteur bancaire

Dans le contexte futur du secteur bancaire, l'accent mis sur les innovations technologiques sera décisif pour attirer les consommateurs et investisseurs. Il est essentiel pour les banques d'investir dans leur numérisation afin d'améliorer l'efficacité opérationnelle en back-office, réduire les risques opérationnels et accroître les profits (Hamilton, 2018). Pour les banques, adopter la technologie numérique est impératif. En effet, la combinaison des demandes des clients qui sont en mutation, la pression pour diminuer les dépenses et pour augmenter l'efficacité, contraint les institutions bancaires à intégrer des technologies de pointe (Hakkaraien, 2022).

De plus, aujourd'hui, des géants du secteur technologique, tels qu'Amazon et Google, se sont aventurés dans le domaine des services financiers. Ils ont rapidement su tirer parti de leurs réseaux étendus, de l'abondance de données clients et de leurs infrastructures technologiques préexistantes pour cibler et proposer des services financiers qui s'intègrent à leurs autres services. Auparavant, les banques étaient seules détentrices des informations financières des clients, ce n'est plus le cas aujourd'hui. En effet, les banques voient désormais leur monopole remis en question par ces acteurs technologiques qui, eux aussi, sont capables d'exploiter leurs données clients pour réaliser des évaluations semblables (Hakkaraien, 2022).

Mistrean (2021), ajoute également que l'objectif principal des activités bancaires est d'augmenter les bénéfices en augmentant les revenus et en renforçant leur présence sur le marché. Cet objectif est atteignable en veillant à la satisfaction des clients des services financiers et bancaires. Afin d'atteindre ses objectifs, il est crucial pour les banques de répondre aux besoins de leurs clients à travers leurs offres de produits et services.

C'est pourquoi, Mistrean (2021), rajoute que l'approche des banques modernes est résolument centrée sur le client, une stratégie mise en place pour améliorer la satisfaction des consommateurs et renforcer la fidélité de ceux qui sont essentiels à leur activité. L'auteur définit l'orientation client comme étant une stratégie qui « signifie garantir un produit ou un service de qualité, à tout moment, pour les clients qui en font la demande » Mistrean (2021).

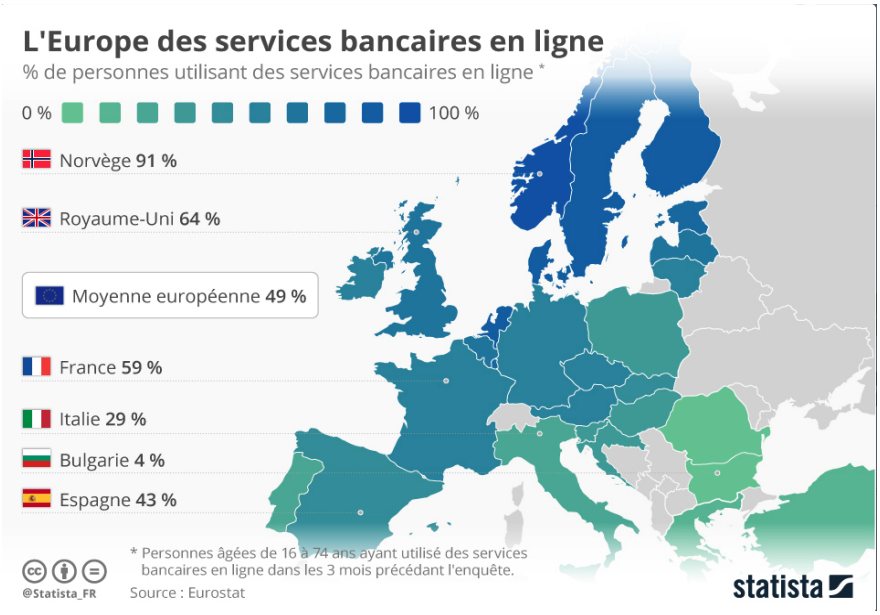
Ainsi, les banques apportent une grande importance à la collecte et l'analyse des informations relatives à leurs clients, qui leur permette d'adapter leurs services aux besoins et aux attentes prioritaires. Cette stratégie orientée vers le client est fondamentale pour assurer le succès et la pérennité de l'établissement financier sur le marché, car elle contribue à la fidélisation de sa clientèle et à l'obtention d'un avantage concurrentiel significatif (Mistrean, 2021).

En revanche, un article de Czímer, Dietz, László, et Sengupta (2022) publié par McKinsey & Company montre que de nombreuses banques traditionnelles sont confrontées à une stagnation, voire même à une diminution de leurs revenus et bénéfices. En effet, le taux de rentabilité sur fonds propres moyen au niveau mondial dans le secteur bancaire atteignait environ 9,5 % en 2021, marquant une nette amélioration par

rapport aux 6 % de 2020, mais cela représente une chute significative par rapport aux 15 % observés avant la crise financière de 2008. Pour 2030, ce taux se voit de diminué et d'atteindre un seuil inférieur à 7,2%.

De plus, la pandémie de COVID-19 a accéléré les banques à se numériser, un processus qui aurait mis plus de temps à se développer sans la crise sanitaire. En effet, avec de nombreuses succursales fermées ainsi qu'une diminution importante des interactions humaines, les banques n'ont pas eu d'autres choix que de numériser une grande partie de leurs services. De plus, la pandémie a également forcé de nombreuses personnes à basculer vers l'utilisation des applications bancaires (Valenti et Alderman, 2022). Un sondage effectuer par Valenti et Alderman (2022), montre qu'un tiers des participants reconnaissent utiliser les plateformes numériques bancaires bien plus souvent qu'avant la crise sanitaire. Ainsi, des banques comme Wells Fargo ont vu les dépôts de chèques augmenter de 35% et une augmentation des virements bancaires de 50%. D'autres établissements bancaires ont observé une augmentation sans précédent de l'emploi de leurs services numériques, attirant ainsi de nouveaux utilisateurs. À titre d'exemple, Bank of America a rapporté que pendant la pandémie, 7 millions de ses clients ont interagi pour la première fois avec Erica, son assistant virtuel.

Figure 1 : L'Europe des services bancaires en ligne : pourcentage de personnes utilisant des services bancaires en ligne dans différents pays européens.



Source: Boittiaux, P. (2017, 15 mars). L'Europe des services bancaires en ligne. Statista. Récupéré de https://fr.statista.com/infographie/8510/leurope-des-services-bancaires-en-ligne/

Selon une enquête menée par Eurostat, 59% des Français ont choisi d'utiliser des services bancaires en ligne, ce qui leur permet d'éviter de se rendre en agence bancaire. Ces services persuadent encore plus les Norvégiens, car 91% d'entre eux les utilisent (Boittiaux, 2017). Cette numérisation a permis aux banques de personnaliser leurs offres et d'être plus proches de leurs clients. C'est pourquoi, en réponse à ces changements, de nombreuses banques augmentent leurs investissements dans leur numérisation afin de maintenir leur singularité et de répondre au mieux aux besoins des consommateurs (Osei, Cherkasova et Oware, 2023).

D’après Gibson (2022), l'abondance et l’accessibilité des données représentent une véritable opportunité pour les banques pour mieux comprendre le comportement, les besoins et les préférences des clients. La compréhension de volume massif permet aux banques de fournir un service plus personnalisé, ce qui permet également de renforcer leur stratégie marketing en offrant des messages personnalisés, en procédant à un meilleur ciblage ainsi qu’en optimisant les offres cross-selling.

Selon un article paru dans la revue "Teller Vision" (2023), Digital Q2 Holdings, Inc., spécialiste des solutions de numérisation pour les secteurs bancaires et de prêt, a récemment diffusé son analyse des tendances et orientations majeures pour les banques de détail en 2023. Cette étude, qui repose sur les réponses de 600 dirigeants du secteur financier provenant de diverses institutions bancaires et unions de crédit à travers le monde, explore les évolutions notables au sein de ces organisations durant 2022 ainsi que leurs axes de concentration pour l'année actuelle.

Quatre tendances dominantes ont été mises en évidence par Q2 pour le secteur bancaire de détail en 2023, dont la suppression des obstacles dans l'expérience client, le développement des offres et des moyens de paiement numériques, l'exploitation des données volumineuses, de l'intelligence artificielle et de l'analytique avancée

De plus, Jim Marous, fondateur et directeur général de Digital Banking Report, (comme cité dans Teller Vision 2023), a exprimé que face aux incertitudes économiques, les établissements financiers devraient se tourner davantage vers l'analyse de données pour devancer les attentes des clients, en générant des interactions favorisant à la fois l'amélioration du bien-être financier des clients et l'optimisation des revenus. Il a insisté sur la nécessité pour les dirigeants bancaires d'adopter de nouveaux paradigmes, allant de la mise en œuvre rapide de solutions numériques à la personnalisation des services grâce à une interaction prédictive.

L’article ajoute que la transition numérique bancaire a aussi initié un nouvel élan de concurrence entre les acteurs financiers traditionnels et les nouveaux venus, mettant les institutions sous pression pour se transformer en organisations prioritaires numériques de manière accélérée.

1.2 Question de recherche

C’est dans ce contexte particulier et dynamique que les banques ont compris qu’il était essentiel de trouver des méthodes innovantes permettant de renforcer et de rendre leur campagne marketing encore plus efficace. Afin d’y parvenir, l’un des éléments cruciaux à prendre en considération se trouve souvent dans la capacité qu’ont les banques à exploiter les données collectées. Cela soulève donc la question de savoir

Comment l'analyse des données peut-elle être utilisée pour cibler efficacement les clients, et personnaliser les offres dans les prochaines campagnes marketing des banques ?

1.3 Objectif de la recherche :

1.3.1 Objectif général :

En se basant sur la problématique ci-dessus, ce travail de recherche vise à étudier les différentes stratégies d'analyse de données en se focalisant sur deux composantes clés : le ciblage des clients et la personnalisation des offres marketing. Cette question de recherche invite à une étude rigoureuse des comportements des clients afin d'offrir un marketing bancaire plus ciblé et personnalisé pour les campagnes futures.

1.3.2 Ciblage des Clients :

Cet objectif a pour but d'analyser les comportements transactionnels des clients ainsi que leurs interactions avec les services bancaires, sur la base des données récoltées. Il s'agit également de déterminer comment l'élaboration de modèles prédictifs, tels que Customer2Vec et les forêts aléatoires, permet d'identifier les segments de clients les plus susceptibles de souscrire à des offres bancaires spécifiques. Cette recherche vise à améliorer la personnalisation des offres et à optimiser les stratégies de ciblage marketing de la banque.

1.3.3 Personnalisation des offres :

Pouvoir, grâce à l'analyse des données, concevoir des stratégies marketing adaptées aux caractéristiques et préférences des clients identifiées permettant d'augmenter l'engagement client et la pertinence des messages.

1.4 Contribution et structure de la recherche :

Ce mémoire vise à approfondir la compréhension de l'analyse de données afin d'améliorer les campagnes marketing futures. De plus, cette recherche a pour objectif de contribuer à la littérature déjà présente sur le marketing bancaire tout en intégrant une analyse pratique basée sur un dataset spécifique. En effet, ce travail espère d'une part enrichir le cadre théorique et d'autre part permettre de délivrer des insights intéressants et applicables pour les professionnels du marketing bancaire.

Enfin, sur base de cette démarche combinant la littérature à la pratique, le travail espère fournir une perspective holistique mettant en lumière l'importance de l'analyse des données pour une prise de décision éclairée et stratégique.

Dans un premier temps, nous allons explorer la littérature qui traite de l'analyse des données ainsi que du marketing bancaire. Cette première étape permettra d'établir le cadre théorique et de mettre en lumière les zones d'ombre dans les connaissances actuelles, des zones que cette thèse s'efforcera d'éclairer.

Dans un deuxième temps, nous allons établir les méthodes utilisées pour la collecte de données ainsi qu'une description du dataset utilisées, des techniques d'analyse statistique, et des méthodes qualitatives appliquées.

Ensuite, une section sera consacrée à la présentation des résultats de l'analyse de données du dataset en mettant en lumière comment ces données peuvent servir à cibler les clients de manière efficace et à personnaliser les offres.

Par la suite, une autre section permettra l'interprétation des résultats ainsi que la mise en commun avec la littérature. Elle abordera également les implications théoriques et pratiques des découvertes.

De plus, un résumé des découvertes sera effectué afin de souligner les contributions de l'étude à la littérature existante et à la pratique professionnelle dans le but de proposer des recommandations pour l'élaboration de futures campagnes marketing bancaires.

Enfin, une dernière section abordera les limites de l'étude actuelle.

2 Revue de la littérature

2.1 Évolutions des données dans le secteur bancaire

L'analyse des données est une pratique qui a connu de nombreuses évolutions depuis ses débuts aux 18e siècle, période durant laquelle les statistiques sont apparues comme une discipline. En effet, Lutz et Lagacherie (2016), expliquent qu'à l'origine, l'analyse des données se basait sur les statistiques avec deux types de techniques bien distinctes :

- La statistique allemande qui se concentre sur des analyses générales, avec un accent particulier sur les aspects qualitatifs.

- La statistique anglaise qui est exclusivement quantitative, autrefois appelée "arithmétique politique" jusqu'en 1798. Cette approche a été pionnière dans la formalisation et l'organisation des études démographiques, en particulier grâce à l'examen détaillé des registres paroissiaux de baptêmes, mariages et décès.

Ce n'est qu'au 20e siècle, avec l'apparition de l'informatique et plus précisément avec la mise en réseau des machines, que les banques ont commencé à se moderniser, transformant des processus purement manuels en processus automatisé (Goetz, 2019).

Cet essor du numérique a non seulement permis d'automatiser les processus, mais a également permis la création d'un univers d'échanges continus de l'information. Ces échanges ont pu donner naissance à un flux important de données marquant l'avènement de l'ère du Big Data dans le secteur bancaire. Pour avoir une idée de l'importance de la génération de l'information, on compte plus d'informations produites en seulement une journée en 2015 que sur toute l'année 1997 (Metge, 2015).

De plus, les données jouent un rôle majeur partout, et le monde de la banque et de la finance ne fait pas exception. Les banques récoltent énormément de données qui, si bien utilisées, peuvent aider à prendre de meilleures décisions, à voir plus clair pour l'avenir et à rendre les clients plus satisfaits. L'utilisation de l'analyse de données, aussi appelée Analytique Bancaire, change la donne pour les banques. Elle les aide

à trouver de nouvelles manières de gagner de l'argent et à rester au top dans un monde de plus en plus numérique (Allied Market Research, 2022).

On s'attend à ce que le secteur de l'analyse de données dans les banques atteigne une valeur de 28,11 milliards de dollars d'ici 2031, ce qui montre bien que de plus en plus de banques commencent à voir son importance. L'analyse de données aide les banques à mieux utiliser leurs informations et est devenue un outil indispensable (Allied Market Research, 2022).

Aujourd'hui le secteur bancaire a en leur possession des données volumineuses, celles-ci souvent désignées sous le terme de Big Data. Dans le cours de "Data Analytics in Marketing" (Session 5, 2023), ICHEC, donnée par Mme Martine George qui se réfère au Chapitre 4 du livre "Data Science: Concepts and Practice" par Vijay Kotu et Bala Deshpande (2019), on retrouve une définition pour le Big Data comme étant "d'énormes volumes de données qui dépassent la capacité de traitement des systèmes de bases de données conventionnels". Ces mégadonnées sont caractérisées par les 3V : Volume, Vitesse et Variété.

- **Volume :** Il représente la quantité importante de données traitées qui selon le type d'information peuvent atteindre les téraoctets et zettaoctet (Dumoulin, 2023).

- **Vitesse :** Elle fait référence à la vitesse à laquelle les données sont générées, collectées et traitées. Dans le contexte du Big Data les données peuvent être générées à une très grande vitesse. Par exemple : le high frequency trading qui représente la vitesse à laquelle les actifs sont échangés sur les différents marchés financiers mondiaux, se fait à l'échelle commune de nanoseconde, soit un milliard de secondes (Bourany, 2019).

- **Variété :** Elle fait référence aux différents types de données que l'on retrouve, elles peuvent être structurées en chiffres et tableaux ou sous forme de descriptions qualitatives avec divers indicateurs. Mais de nos jours, une grande majorité des données, soit 80%, sont en réalité non structurées. On trouve parmi elles des textes simples, des séquences d'images, de vidéos, ou même des informations complexes issues de capteurs ou de séquençages d'ADN, qui introduisent des défis inattendus. Les méthodes d'analyse statistique traditionnelles, conçues pour des données numériques standard, ne sont pas adaptées pour déchiffrer des informations aussi complexes. Heureusement, les techniques modernes d'apprentissage automatique offrent maintenant des outils pour traiter et interpréter ce type de données diversifiées (Bourany, 2019).

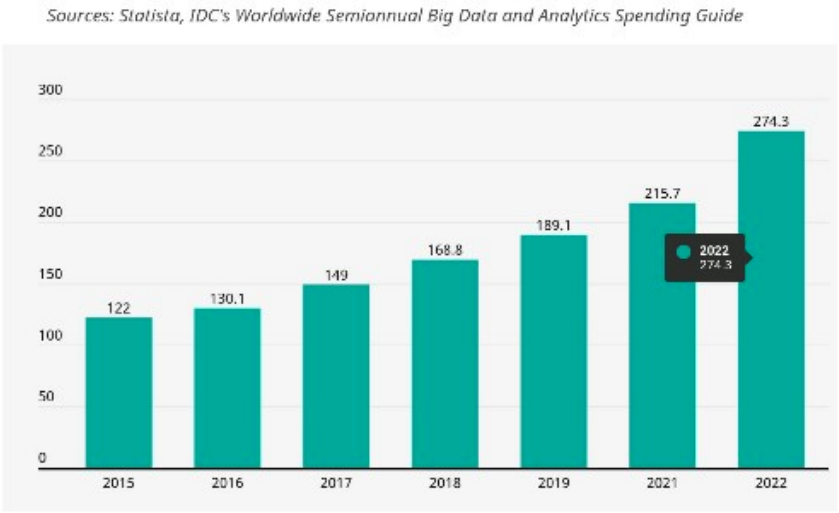
Ces mines de données nécessitent d'être traitées correctement afin de permettre une analyse précise et approfondie. En effet, selon Buzullier (2019), "Si l'on ne sait pas traiter correctement la donnée, celle-ci perd toute sa valeur et fausse la prise de décision". C'est dans ce contexte qu'un nouveau terme émerge et gagne en importance : la Smart Data.

Le Smart Data signifie en français "données intelligentes" ces données regroupent les informations pertinentes aidant les entreprises dans leurs processus de prise de décision. Il est souvent vu comme étant l'évolution logique du Big Data qui lui se rattache principalement à la récolte de données souvent dans le désordre. Grâce à l'aide d'algorithmes spécifiques, les données récoltées peuvent être triées, permettant à la smart data d'avoir une utilité supérieure à celle du Big Data. Ces deux types de données se complètent mutuellement. En effet, dans une perspective d'amélioration de la gestion de la relation client, une entreprise trouverait avantageux d'incorporer ces deux stratégies dans ses processus opérationnels (Brossault, 2023).

Le Big Data était traditionnellement rattaché au 3V (Volume, Vitesse, Variété). Cependant, le Smart Data va au-delà de cette approche en ajoutant 2V supplémentaire : Valeur et Véracité

- **Valeur :** “la capacité de ces données à générer du profit” (Bourany, 2019).
- **Véracité** : “leur validité, c.-à-d. qualité et précision ainsi que leur fiabilité” (Bourany, 2019).

Figure 2 : Revenus du big data et de l'analyse d'entreprise dans le monde de 2015 à 2022 (en milliards de dollars américains).



Source: Revenus du big data et de l'analyse d'entreprise dans le monde de 2015 à 2022 (en milliards de dollars américains). Adaptée de Statista et IDC's Worldwide Semiannual Big Data and Analytics Spending Guide, cité dans Coret, S. (2021, 27 septembre). Les dépenses mondiales en matière de big data et d'analyse d'entreprise.

Le graphique présenté illustre l'évolution des revenus générés par le big data et l'analyse d'entreprise à l'échelle mondiale entre 2015 et 2022, exprimés en milliards de dollars américains. On constate une progression significative au fil des années. En 2015, l'industrie commençait avec un revenu de 122 milliards de dollars. Cet élan s'est maintenu et même accéléré, atteignant les 274,3 milliards en 2022. On constate que les données qui sont traitées correctement peuvent générer un revenu très important pour les entreprises.

Enfin, la véracité de ces informations est devenue une préoccupation majeure, surtout avec l’accroissement des fake news provenant des réseaux sociaux. Il est crucial que les données soient véridiques, de qualité et fiables pour qu'elles soient vraiment utiles. Cette inquiétude est d'autant plus forte quand il y a un manque surprenant d'informations sur d'où viennent ces données et comment elles sont collectées (Bourany, 2019).

2.2 Marketing bancaire

2.2.1 Définitions :

Avant de se lancer dans l'exploration du marketing bancaire, il est essentiel de saisir les deux idées fondamentales qui en constituent la fondation. Ces concepts ne sont pas juste des termes techniques, ils représentent les piliers sur lesquels repose toute l'approche marketing dans ce secteur si particulier.

Le premier concept à définir est celui du marketing, il existe énormément de définitions qui décrivent ce qu'est le marketing. Selon Philip Kolter (2012), le marketing est "la science et l'art d'explorer, de créer et de fournir de la valeur pour satisfaire les besoins d'un marché cible avec un profit". Cette définition met en lumière la création de valeur ainsi que la recherche permanente de satisfaction des besoins des clients.

Le deuxième concept est celui du marketing bancaire Hévin (2023), qui offre une définition éclairante du marketing bancaire, le décrivant comme "l'ensemble des méthodes et techniques utilisées par les banques et autres institutions financières pour promouvoir leurs produits et services. Ces institutions cherchent à attirer de nouveaux clients, fidéliser leur clientèle existante et augmenter leur part de marché. Grâce au marketing bancaire, les banques sont en mesure de mieux comprendre les attentes de leurs clients et ainsi proposer des offres personnalisées." Cette définition met en lumière non seulement les objectifs principaux du marketing bancaire, mais aussi l'importance capitale de comprendre les besoins et attentes des clients pour le succès des institutions financières.

2.2.2 Fondement théorique

Le marketing bancaire est passé de la publicité traditionnelle et de la vente personnelle à une approche plus intégrée incluant le développement de produits, la tarification, la distribution et le marketing numérique. Les banques se concentrent désormais sur l'identification et la satisfaction des besoins des clients de manière plus efficace et efficiente et essayent de se démarquer de leurs concurrents, en mettant l'accent sur la création de valeur et la mise en place de relations à long terme (Mihaela, 2013).

Les fondements théoriques du marketing bancaires regroupent une série de techniques utilisées par les banques pour promouvoir leurs produits et services afin d'acquérir et fidéliser les clients, mais aussi pour augmenter leur part de marché. Ces techniques reposent principalement sur l'analyse de données clients, la segmentation de la clientèle ainsi que sur le positionnement stratégique entre tradition et innovation. Le rôle du digital joué dans la stratégie de marketing bancaire, comme l'utilisation des réseaux sociaux, des applications mobiles et des sites web, montre l'évolution, du secteur qui ne cesse d'accroître (Mihaela, 2013).

De plus, l'apparition du business intelligence a permis aux banques de voir leurs interactions avec leurs clients ainsi que les optimisations de leurs services complètement changer. Audzeyeva et Hudson (2016), soulignent l'importance de l'application du business intelligence (BI) dans le secteur bancaire, affirmant que l'exploitation stratégique des données peut transformer la stratégie marketing d'une banque. L'intégration de la BI permet une compréhension profonde des besoins des clients, ce qui est crucial pour le développement de produits et services bancaires adaptés.

Une définition intéressante donnée par Muntean (2018), qui définit la Business Intelligence (BI) comme “un terme générique désignant les stratégies, les technologies et les systèmes d'information utilisés par les entreprises pour extraire, à partir de données volumineuses et diverses, selon la chaîne de valeur, des connaissances pertinentes pour prendre en charge un large éventail d'activités opérationnelles, tactiques et stratégiques”.

2.2.3 Segmentation de la clientèle :

La segmentation du marché de consommation est cruciale en marketing stratégique, car elle influence directement la compétitivité et la rentabilité de l'entreprise. Un segment clairement défini est essentiel. L'objectif principal de cette segmentation n'est pas seulement de repérer des groupes spécifiques ayant des caractéristiques distinctes sur le marché, mais aussi de trouver des groupes qui expriment des besoins spécifiques pour un produit ou un service très différencié (Rozhko, 2023).

Sahrir (2024), définit la segmentation comme un processus qui « implique de regrouper des marchés divers en consommateurs potentiels ayant des besoins ou des caractéristiques similaires, qui réagissent de manière similaire lorsqu'ils font leurs achats. »

Étant donné la diversité du marché, les producteurs rencontrent des difficultés pour y répondre de manière exhaustive. Par conséquent, les spécialistes du marketing doivent choisir de se concentrer sur des segments spécifiques, généralement homogènes, qui correspondent à la capacité de l'entreprise à satisfaire la demande. En classant les clients en segments spécifiques, les banques peuvent adapter leurs stratégies marketing au sein du système bancaire, et la segmentation sert également de base pour déterminer les marchés cibles et les emplacements, ce qui permet aux banques d'ajuster leurs services aux besoins des clients et d'améliorer l'efficacité de la prise de décision. Par conséquent, la segmentation est une décision judicieuse. Les banques doivent également prendre en compte d'autres facteurs tels que les caractéristiques personnelles et les antécédents des clients, qui influent sur leurs décisions et leurs évaluations des produits bancaires (Sahrir, 2024).

Mousaerid (2020), définit la segmentation comme étant un processus qui « consiste à répartir les clients en groupes présentant certaines caractéristiques communes. » De plus, il affirme qu’une segmentation bien élaborée permet une allocation optimale des ressources marketing comme les systèmes de recommandation dans le secteur bancaire. En effet, cela permet d’identifier les groupes de clients les plus aptes à répondre positivement à une offre spécifique.

Une étude réalisée par Osei, Ampomah, Kankam-Kwarteng, Bediako et Mensah (2021), souligne l’importance de la corrélation positive entre la segmentation et la satisfaction des clients dans le secteur bancaire.

L'enquête a concerné 200 participants provenant de cinq banques situées dans la grande ville de Kumasi, au Ghana. Concernant la répartition démographique, elle incluait 67 % d'hommes et 33 % de femmes. Pour ce qui est de l'âge des répondants, 39,5 % avaient entre 18 et 30 ans, 26,5 % entre 31 et 40 ans et 34 % se situaient dans la tranche d'âge de 41 à 50 ans.

Types de Segmentation et Satisfaction des Clients :

- **Segmentation géographique :** L'étude a révélé un lien positif entre la segmentation géographique et la satisfaction des clients, soulignant l'importance d'ajuster les offres et services bancaires aux exigences spécifiques des clients résidant dans diverses régions géographiques.

- **Segmentation démographique :** Une corrélation positive a également été observée avec la satisfaction des clients, ce qui illustre l'utilité de viser les clients en fonction de critères démographiques tels que l'âge, le sexe ou le revenu.

- **Segmentation comportementale :** L'enquête a mis en évidence une influence positive de la segmentation comportementale sur la satisfaction des clients, mettant en avant l'intérêt de saisir et d'adapter les services aux habitudes comportementales des utilisateurs des services bancaires.

Les conclusions de l'étude ont confirmé l'impact positif de la segmentation géographique, démographique et comportementale sur la satisfaction client dans le domaine bancaire.

En détail, les coefficients des chemins analysant les relations entre la segmentation démographique et la satisfaction client étaient de 0,228 entre la segmentation géographique et la satisfaction client de 0,520, et entre la segmentation comportementale et la satisfaction client de 0,267, tous statistiquement significatifs avec un $p < 0,05$.

La valeur P représente la probabilité qu'une différence observée entre les groupes soit le résultat du hasard, sous l'hypothèse qu'il n'y a aucun effet ou aucune différence réelle (hypothèse nulle). Elle mesure ainsi à quel point il est improbable que la différence observée soit due au hasard, avec des valeurs proches de zéro indiquant une forte probabilité que la différence soit réelle. Les valeurs de P près de 1 suggèrent que la différence entre les groupes est probablement due au hasard. Les termes tels que "très significatif" ou "fortement significatif" sont souvent utilisés dans les publications médicales pour indiquer la proximité de la valeur P avec zéro (Dahiru, 2011).

Au vu de ces résultats, l'étude a suggéré aux établissements bancaires d'élaborer et d'implémenter une stratégie intégrant des critères géographiques, démographiques et comportementaux pour perfectionner la qualité de service et augmenter la satisfaction de la clientèle.

2.2.4 Offre personnalisée :

Czímer et al. (2022), affirment que les forces économiques et technologiques rendent le modèle traditionnel des banques dépassé. En effet, ils rajoutent que les investisseurs sont de plus en plus à la recherche de spécialisation radicale. Avec l'émergence des plateformes intersectorielles, les banques se trouvent désormais dans le besoin de rivaliser avec à n'importe quelle entité prête et capable de proposer divers types de services financiers. Des acteurs majeurs du secteur technologique, tels que Google et Tencent, ont déjà intégré des services bancaires à leurs offres, les rendant accessibles à des millions d'utilisateurs sans effort. De plus, on assiste à une explosion de nouveaux venus dans le secteur un peu partout dans le monde.

Ensuite, selon eux, la révolution technologique a rendu obsolète l'idée qu'une grande taille est synonyme de meilleurs services, de loyauté, de client accru, ou d'une capacité supérieure à collecter et analyser les

données. Depuis 2015, pas moins de 200 banques numériques ont vu le jour, dans le domaine de la banque d'investissement tel que les sociétés d'acquisition spécifiques, ainsi que des plateformes de paiement et de financement de startups comme SeedInvest Technology (Czímer et al. 2022).

Revolut est un exemple de néobanque, les néobanques, parfois appelées « banques challengers », sont des entreprises de technologie financière qui offrent des applications, des logiciels et d'autres technologies pour simplifier la banque mobile et en ligne. Ces fintechs se spécialisent généralement dans des produits financiers particuliers, comme les comptes chèques et les comptes d'épargne (Walden, 2021).

Elles offrent une grande variété de personnalisation pour ces clients en proposant aux utilisateurs une flexibilité selon leurs besoins spécifiques. Avec des options allant de la carte standard gratuite à la carte Ultra haut de gamme, Revolut répond à une large gamme de besoins financiers et de styles de vie. Une des offres que Revolut propose est la carte Revolut Ultra qui répond aux besoins d'une clientèle premium avec des avantages comme des accès illimités aux salons d'aéroport, une annulation possible, peu importe la raison, un accès au Financial Time, Wework ainsi que NordVPN. Des opérations boursières sans commission et pleines d'autres avantages (Revolut, s.d)

Un autre exemple illustré par Czímer et al. (2022), est celui de la Banque Royale du Canada (RBC) qui a répondu aux nombreux défis bancaires en proposant des offres diverses et variées. Comme la RBC Ventures qui est une initiative axée sur le financement de nouvelles entreprises et la formation de partenariats stratégiques, elle a réussi à toucher 3,2 millions de Canadiens.

En outre, des innovations telles que Ownr qui se présente comme une solution tout-en-un pour les petites entreprises et startups, offrant des services allant de la conception de sites internet à l'enregistrement d'entreprises et aux prestations bancaires, le tout sur une unique plateforme. Ou encore GarbageDay une application utilisée par près de 200 000 canadiens qui aide les citoyens à ne pas oublier les jours de ramassage des déchets et du recyclage, les obligations saisonnières, et fournit des informations sur le stationnement. Grâce à cela, la RBC a réussi à attirer des milliers de nouveaux clients (Czímer et al., 2022).

Cette vague de concurrents a considérablement augmenté les attentes des consommateurs. Aujourd'hui, les clients, qu'ils agissent à titre individuel ou au sein d'organisations, exigent une vaste gamme de services de la part de leurs fournisseurs de services financiers. Les études montrent qu'ils privilégient notamment un degré élevé de personnalisation dans les services qui leur sont proposés (Czímer et al., 2022).

Enfin, Hakkaraien (2022), rajoute que l'attrait et la facilité d'utilisation semblent être les principaux facteurs du succès des nouveaux venus sur le marché. Autrement dit, la clientèle valorise la capacité d'accéder à l'ensemble de leurs services bancaires en ligne ou par le biais d'appareils mobiles. Plus l'utilisation du service est aisée, mieux c'est. De plus, la clientèle a la possibilité de choisir parmi une gamme d'offres plus adaptées à leurs besoins, y compris des services qui s'étendent au-delà du financier. Ces offres sont conçues sur la base d'importantes quantités de données recueillies par les fournisseurs concernant les activités quotidiennes de leurs clients.

2.3 L'analyse prédictive dans le secteur bancaire :

L'analyse prédictive émerge comme une clé essentielle pour déchiffrer l'avenir et orienter les décisions stratégiques. Elle représente une branche sophistiquée de l'analytique, utilisée pour anticiper

des événements futurs inconnus. Kumari et Aggrawal (2022), définissent l'analyse prédictive comme étant « une branche de l'analyse avancée utilisée pour faire des prédictions sur des événements futurs inconnus. Il déploie de nombreuses techniques telles que l'exploration de données, l'intelligence artificielle et l'apprentissage automatique pour analyser les données actuelles afin de faire des prédictions futures. »

Dans le cadre du cours « Data Analytics in Marketing » (Session 5, 2023), ICHEC, Mme Martine George qui se réfère au Chapitre 4 du livre « Data Science: Concepts and Practice » par Vijay Kotu et Bala Deshpande (2019), définit un modèle prédictif comme « un algorithme mathématique qui prédit une variable cible à partir d'un certain nombre de variables explicatives. »

Pour les banques, cette technique s'avère être un atout majeur, en effet elle permet une compréhension plus approfondie sur les besoins des clients leur permettant de lancer des produits et services innovants basés sur la recherche client, la segmentions et l'analyse des données. Elle permet également de fournir des expériences personnalisées et innovantes (Cognizant, s.d).

La plupart des problèmes d'analyse prédictive peuvent être regroupés en deux catégories distinctes : classification ou prédiction de valeurs numériques. Lors de la classification ou de la détermination de catégories, l'objectif est d'exploiter les données fournies par les variables prédictives ou indépendantes afin de répartir les échantillons de données dans deux catégories distinctes ou plus. Pour ce qui est de la prédiction de valeurs numériques, le but est de déduire la valeur numérique de la variable dépendante à partir des valeurs adoptées par les variables indépendantes (George, M, 2023).

Dans le domaine de la classification et de l'évaluation des données, différentes méthodes sont mises en œuvre pour classer et évaluer les diverses structures de données. On distingue principalement deux catégories de données selon leur structure : les données structurées et les données non structurées. Les données structurées sont généralement conservées dans des bases de données et incluent des types de données spécifiques ainsi que des désignations de champs. Cette organisation structurée facilite la classification et l'évaluation précises des données, nécessitant une classification rigoureuse. Par ailleurs, cette démarche exige de classifier et d'évaluer les résultats pour chaque colonne de manière individuelle.

En contraste, les données non structurées se trouvent souvent sous forme de fichiers logs ou de documents, contenant des significations contextuelles. L'utilisation des techniques de Traitement du Langage Naturel (TLN) permet une analyse sémantique qui révèle des informations sensibles dissimulées dans les textes. Dans ce cadre, le degré de précision de la classification peut aller du général au spécifique. Une classification générale fournit des résultats pour le document dans son ensemble, tandis qu'une classification détaillée nécessite de repérer des types spécifiques d'informations sensibles présentes dans le document et d'indiquer leurs niveaux de sensibilité respectifs (Zu, QI, Li, Men, Lu, Ye et Zhang, 2024).

Selon le document de l'University of Illinois Board of Trustees sur la Data Science and Predictive Analytics (s. d.), les méthodes courantes de modélisation prédictive les plus populaires sont :

- **Réseaux de neurones :** Ce sont des méthodes avancées capables de capturer des liaisons extrêmement complexes entre données. Ce modèle doit sa popularité grâce à sa capacité à pouvoir traiter des dynamiques non linéaires présentes dans les données volumineuses. Les réseaux de neurones complètent souvent les analyses réalisées par des techniques plus simples, telles que la régression et les arbres de décision, en offrant une modélisation basée sur la

reconnaissance de motifs et l'imitation des processus cérébraux humains. Ils représentent une frontière innovante dans le domaine de la modélisation prédictive.

- **L'Arbre de décision** : Ces modèles catégorisent les données en différents sous-groupes basés sur les catégories des variables d'entrée, facilitant la compréhension des séquences décisionnelles. Ils sont largement reconnus pour leur approche intuitive.

- **Régression :** Une technique statistique très répandue, l'analyse de régression vise à déterminer les relations entre différentes variables. Elle est particulièrement adaptée aux données continues susceptibles de se conformer à une distribution normale et est efficace pour identifier des motifs importants au sein de vastes ensembles de données.

2.3.1 Customer2vec

Dans le secteur bancaire, une méthode avancée appelée Customer2Vec utilise une combinaison de techniques de classification neuronale et de clustering pour améliorer la segmentation des clients et identifier les points communs entre les clients, améliorant ainsi la qualité de la segmentation. Customer2Vec, s'inspire du succès de Word2Vec dans le domaine du traitement du langage naturel. Le principe du Word2Vec est d'attribuer à chaque mot un vecteur qui représente le sens du mot et la similarité entre différents mots dans un document. Une fois cette étape effectuer le modèle va à chaque itération ajuster ces vecteurs en fonction de la proximité des mots dans le texte (Mousaeirad, 2020).

En analysant les empreintes digitales des clients, également appelées vecteurs, il devient possible de les catégoriser de manière plus intuitive, en regroupant des individus ayant des intérêts ou des habitudes d'achat similaires. Dans le contexte bancaire, cela pourrait impliquer d'identifier les clients susceptibles d'être intéressés par un prêt ou un compte d'épargne particulier, en tenant compte non seulement de leurs interactions passées, mais également en procédant à une analyse approfondie de leurs comportements et caractéristiques communs (Mousaeirad, 2020).

L'idée est simple : au lieu de se casser la tête à choisir et à créer des caractéristiques une par une, on laisse un réseau de neurones s'occuper de tout. Ce réseau va prendre toutes les infos et nous donner en retour une sorte de résumé serré sur chaque utilisateur, qu'on appelle un "embedding". Avec ce résumé, on peut facilement voir qui ressemble à qui, que ce soit des utilisateurs ou des produits. Et le plus souvent, ce petit résumé fait par la machine est bien plus efficace pour comprendre nos données que si on avait tout fait à la main, surtout si le volume de données à traiter est important (Erokhin, 2022).

Figure 3 : Étape du modèle Customer2Vec

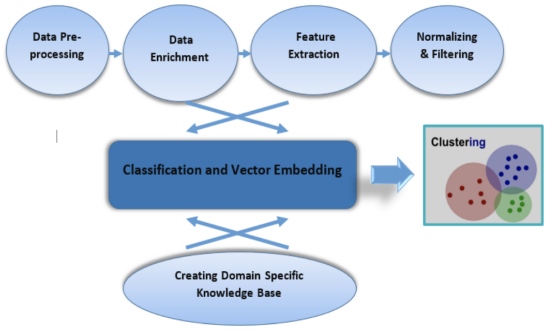


Figure 2. An overview of Customer2Vec Model.

Source: Mousaeirad, S. (2020). Intelligent Vector-based Customer Segmentation in the Banking Industry. *ArXiv*, https://doi.org/10.48550/arXiv.2012.11876

La figure 3 représente graphiquement les diverses phases du modèle, dont voici les explications fournies par Mousaeirad (2020).

- **Data Pre-processing:** La première étape consiste à nettoyer les données et les préparer. Cette étape inclut la suppression de certaines données du client, la correction des erreurs ainsi que la standardisation des formats de données.

- **Data Enrichement**: Cette étape consiste à enrichir les données avec des données dérivées des données existantes ou bien qui proviennent d'autres sources. Le but de cette étape est d'avoir une qualité de données supérieures afin d'avoir une meilleure analyse.

- **Feature Extraction :** Ici les caractéristiques significatives sont identifiées et extraites de l'ensemble des données. La raison de cette étape est de choisir les données les plus pertinentes pour la segmentation des clients.

- **Normalizing & Filtering** : Une fois les caractéristiques identifiées et extraites, elles sont normalisées, c'est-à-dire, elles sont ramenées à une échelle commune afin que certaines caractéristiques ne dominent pas les autres en raison de la différence d'échelle, ce qui peut fausser les résultats des algorithmes d'analyse. Les données sont ensuite filtrées afin de ne sélectionner uniquement les caractéristiques les plus utiles.

- **Classification and Vector Embedding** : Les clients sont ensuite classés et leurs caractéristiques sont transformées en vecteurs dans un espace vectoriel. Cela aide à représenter les clients dans un format qui peuvent être facilement traités par des algorithmes d'apprentissage automatique.

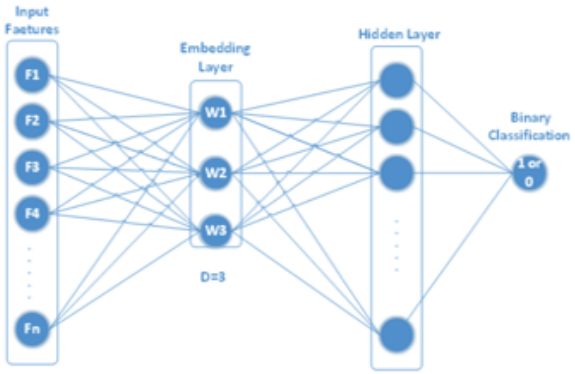
- **Creating Domain Specific Knowledge Base :** Utilisation des connaissances spécifiques au secteur pour interpréter les vecteurs de caractéristiques et les résultats de la classification.

- **Clustering** : En utilisant les vecteurs de caractéristiques, les clients sont regroupés en segments basés sur leurs similitudes. Les groupes résultants peuvent aider à cibler les ressources marketing et à personnaliser les services pour différents types de clients.

Mousaeirad (2020), a utilisé ce modèle dans sa recherche dont le but est de classifier les clients afin de prédire leur risque de défaut de prêt. L'auteur extrait les vecteurs clients et segmente les clients en fonction du risque de crédit du client.

Mousaeirad (2020), explique dans sa recherche que le but du Customer2Vec vise à générer pour un client un code numérique qui permettra aux analystes de les organiser visuellement en différents groupes. Ainsi le modèle vise à intégrer au sein du réseau neuronal une couche de représentation vectorielle, dotée de trois nœuds (W1, W2, W3), qui produit un vecteur en trois dimensions. Ensuite sur cette couche une autre couche cachée supplémentaire est ajoutée qui précède un classificateur binaire. Ce dernier réalise la segmentation en se fondant sur l'attribut désiré. Dans le cas de l'étude, où la segmentation des clients de banque se fait sur la base de leur risque de défaut de paiement, le classificateur évalue si un client, représenté par un vecteur spécifique obtenu des données d'apprentissage, présente un antécédent de défaut de paiement. L'objectif est d'évaluer la probabilité qu'un client ne rembourse pas son prêt.

Figure 4 : Architecture d'un réseau de neurones pour la classification binaire et l'embedding de vecteurs clients



Source: Mousaeirad, S. (2020). Intelligent Vector-based Customer Segmentation in the Banking Industry. *ArXiv*, https://doi.org/10.48550/arXiv.2012.11876

La Figure 4 illustre le réseau de neurones utilisé pour la classification binaire. Pour créer un tel code en trois dimensions, Mousaeirad (2020), utilise une structure spéciale dans le programme d'ordinateur appelé "réseau de neurones", qui comprend une section pour compresser les informations (Embedding Layer), suivies d'une autre section pour affiner l'analyse (Hidden layer), et finalement une partie qui prend une décision "oui ou non" basée sur ce que nous cherchons à comprendre (Binary classification). À la fin, il nous donne un code en trois chiffres pour chaque client, qui résume son profil de risque de défaut de prêt.

- **Input Features** : Les données initiales, telles que des informations sur les clients, sont introduites dans le modèle.

- **Embedding Layer** : Les données sont transformées en vecteurs tridimensionnels pour résumer les informations (W1, W2, W3).

- **Hidden Layer** : Le vecteur est ensuite traité pour détecter des motifs plus complexes.

- **Classification binaire** : Le modèle prend une décision finale, classant l'entrée comme 1 ou 0.

- **Processus itératif** : Le modèle ajuste ses paramètres internes pour améliorer ses prédictions au fil des entraînements.

Ensuite vient le clustering, dans cette étape Mouaeirad (2020), se concentre sur le "clustering" ou regroupement des vecteurs clients, qui a été effectué après avoir entraîné un réseau neuronal entièrement connecté et optimisé les poids pour produire des vecteurs tridimensionnels. Ce processus vise à créer une visualisation des segments de clients dans un système de coordonnées tridimensionnel. Quatre méthodes de clustering ont été utilisées : K-means, Mean-Shift, Mélange de Gaussiennes (Gaussian Mixture) et Carte auto-organisatrice (SOM).

Le modèle Mean-Shift est capable de déterminer le nombre optimal de clusters par lui-même, tandis que pour les autres méthodes, divers nombres de clusters de 2 à 6 ont été testés. Pour évaluer le nombre optimal de clusters, la méthode du "Knee" (le coude) a été utilisée, qui cherche le point de rupture dans le graphique de la somme des erreurs au carré (SSE). Cependant, le choix du nombre de clusters doit aussi tenir compte des objectifs de segmentation et des avis d'experts.

De plus, les clusters créés à partir de ces vecteurs ont été évalués en utilisant trois indices : le score Silhouette, l'indice Calinski-Harabasz (CH), et l'indice Davies-Bouldin. Ces indices aident à déterminer la qualité des clusters, en mesurant la cohérence interne des clusters et la séparation entre eux (Mousaeirad, 2020).

Les résultats du clustering avec les quatre méthodes montrent que Mean-Shift a sélectionné 3 clusters pour l'expérience. K-means, Gaussian mixture et SOM ont été expérimentés avec des nombres de clusters allant de 2 à 6, et les résultats sont présentés dans un tableau. Il semble que K-means a légèrement de meilleurs résultats, et en termes de performance et de scalabilité, toutes les méthodes étaient comparables (Mousaeirad, 2020).

Mousaeirad (2020), résume les résultats du modèle proposé en termes de précision de la classification, du clustering et de la visualisation des vecteurs incorporés. L'un des points forts du modèle est la possibilité de réaliser une segmentation subjective des clients, permettant à l'analyste bancaire de choisir des caractéristiques spécifiques pour la segmentation. De plus, l'inclusion des traits de personnalité dans l'analyse a montré une influence sur les niveaux de risque des clients et a amélioré la précision de la prédiction du défaut de remboursement des prêts. Les vecteurs intégrés aident également à détecter les similitudes entre les clients, ce qui peut influencer les politiques et services proposés aux clients.

2.3.2 L’arbre de décision :

L'arbre de décision est un modèle prédictif très utilisé dans l'analyse de données pour classifier et prédire des résultats basés sur des observations. Un arbre de décision est comparable à un schéma qui guide les décisions, ou un arbre renversé, où chaque nœud examine une caractéristique. À l'extrémité de l'arbre, il y a ce qu'on appelle un nœud terminal qui propose une prédiction relative à la variable d'intérêt, selon les critères établis tout au long du chemin suivi dans l'arbre. Chaque nœud travaille à subdiviser

l'ensemble des données en portions plus homogènes. L'objectif de cette division est d'obtenir des groupes de données les plus uniformes possibles. Prenons l'exemple de deux indicateurs, l'âge et le poids, pour prédire si quelqu'un va s'abonner à un club de fitness. Si les données d'apprentissage montrent que 90 % des individus de plus de 40 ans se sont abonnés, alors on pourrait diviser les données en deux groupes distincts : ceux de plus de 40 ans et ceux de moins. Le groupe des plus de 40 ans serait alors considéré comme étant « pur à 90 % » par rapport à la catégorie d'abonnement (Kotu et Deshpande, 2019).

Pour illustrer ce concept, nous utiliserons l'exemple du jeu de données de golf classique, tel qu'il est présenté dans le cours de Data Analytics In Marketing de George (2023). Dans cet exemple, l'objectif est de prédire si les golfeurs préféreront jouer ou non selon les conditions météorologiques données. Ces conditions sont résumées par des variables telles que l'ensoleillement, la température, l'humidité et le vent. Les historiques de ces conditions, associés aux préférences des joueurs, permettent de construire un modèle d'arbre de décision.

Un arbre de décision divise les données en sous-ensembles plus petits et plus précis en utilisant des règles de décision simples, dérivées des données d'entraînement. Par exemple, si l'on considère les préférences de jeu de golf basées sur l'ensemble de données fourni, l'arbre pourrait poser une série de questions pour aboutir à une décision finale : "Est-ce ensoleillé ?" Si oui, "Est-il trop venteux ?" En fonction des réponses à ces questions, l'arbre de décision mène à une conclusion, soit jouer, soit ne pas jouer.

Dans le cas de l'ensemble de données de golf, nous avons des exemples où des jours spécifiques avec des conditions de température, d'humidité et de vent sont classés soit comme des jours où les golfeurs joueront (Yes), soit comme des jours où ils ne joueront pas (No). L'arbre utilise ces informations pour apprendre quelles caractéristiques et quelles valeurs de ces caractéristiques sont les plus indicatives de la décision de jouer.

Tableau 1 *:* Conditions météorologiques et de la décision de jouer au golf

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	FALSE	no
sunny	80	90	TRUE	no
overcast	83	78	FALSE	yes
rain	70	96	FALSE	yes
rain	68	80	FALSE	yes
rain	65	70	TRUE	no
overcast	64	65	TRUE	yes
sunny	72	95	FALSE	no
sunny	69	70	FALSE	yes
rain	75	80	FALSE	yes
sunny	75	70	TRUE	yes
overcast	72	90	TRUE	yes
overcast	81	75	FALSE	yes
rain	71	80	TRUE	no

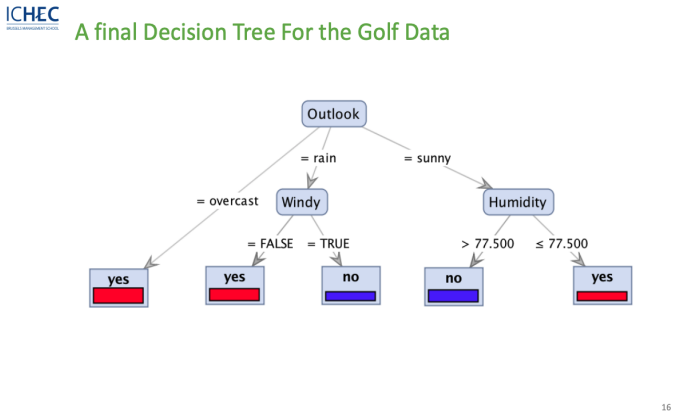
Source: George, M. (2023). Data Analytics in Marketing - Session 7: Performance & Evaluation. ICHEC.

Pour construire un arbre de décision à partir de cet ensemble de données, nous commençons par déterminer le meilleur attribut à utiliser comme nœud de décision. Cette décision est prise en évaluant la capacité de chaque attribut à séparer les données en groupes homogènes en termes de résultat (jouer ou

ne pas jouer). Les attributs sont testés à l'aide de mesures statistiques comme le gain d'information ou l'indice Gini pour déterminer lequel fournit la meilleure séparation.

Après avoir construit l'arbre, nous pouvons l'utiliser pour prédire si un joueur jouera ou non en fonction des conditions météorologiques. Par exemple, pour un jour donné où l'ensoleillement est prévu, l'humidité est en dessous de 77,5 et il n'est pas venteux, l'arbre prédit que le joueur jouera (Yes). Ce processus est répété pour chaque nouvelle observation ou jour pour lequel nous voulons prédire la décision de jouer.

Figure 5 : Arbre de décision



Source: George, M. (2023). Data Analytics in Marketing - Session 7: Performance & Evaluation. ICHEC.

Les arbres de décision offrent l'avantage d'être faciles à comprendre et à interpréter, car ils peuvent être visualisés graphiquement et ressemblent à une suite de règles de décision. En outre, en évaluant les performances du modèle sur un ensemble de données de test, comme l'ensemble de données "out-of-bag" dans notre exemple, nous pouvons obtenir une estimation de la précision de notre modèle. Dans l'exemple de golf, le modèle a atteint une précision d'environ 64%, en ayant correctement prédit 9 des 14 décisions de jouer.

2.3.2.1 Points forts :

Gupta, Arora, Rawat,Jain et Dhami (2017), ont attribué plusieurs avantages à l'arbre de décision comme le fait qu'ils offrent une visualisation claire, rendant leur compréhension et interprétation accessible. Ils demandent également peu de prétraitement des données, contrairement à d'autres méthodologies qui requièrent souvent de normaliser les données, de générer des variables et d'éliminer les valeurs manquantes. De plus, le coût d'exploitation de l'arbre (pour effectuer des prédictions) est proportionnel au logarithme du nombre de données utilisées pour former l'arbre.

Ensuite, ils rajoutent que l'arbre de décision est capable de traiter à la fois des données quantitatives et qualitatives, ainsi les arbres de décision se distinguent d'autres approches qui sont généralement conçues pour un type spécifique de données.

Pour terminer Gupta et al. (2017), affirment qu'ils sont aptes à résoudre des problématiques avec plusieurs issues possibles et que l'arbre se base sur un modèle explicite, où les décisions prises peuvent être

explicitées par des règles logiques simples, offrant typiquement des réponses binaires telle que "oui" ou "non".

2.3.2.2 Limites :

L'une des limites de cet algorithme, selon Molnar (2023), est le manque de fluidité dans l'algorithme. En effet, il explique que des variations mineures dans les données d'entrée peuvent entraîner des prédictions très différentes, ce qui n'est généralement pas le résultat souhaité. Molnar(2023), donne un exemple afin d'illustre ce phénomène. Il prend l'exemple d'un arbre destiné à estimer le prix d'une propriété en se basant, entre autres, sur sa superficie. Si une découpe est fixée à 100,5 m², un écart minime dans la mesure de la surface peut influencer de manière disproportionnée l'estimation du prix. Par exemple, un calcul initial basé sur 99 m² pourrait donner une estimation de 200 000 €, mais l'ajout d'une pièce de 2 m², suite à une nouvelle mesure, pourrait faire grimper cette estimation à 205 000 € avec une saisie de 101 m², malgré une différence de surface minime.

De plus, les arbres de décision peuvent s'avérer instables. Des changements mineurs dans les données peuvent conduire à la construction d'un arbre radicalement différent, du fait que chaque division est conditionnée par les divisions qui l'ont précédée. Un tel changement de structure avec de petites modifications dans les données peut affecter la confiance accordée au modèle (Molnar, 2023).

En outre, Molnar (2023), ajoute que bien que les arbres de décision soient reconnus pour leur interprétabilité, cela reste vrai uniquement pour les arbres de petite taille. La complexité et le nombre de feuilles augmentent exponentiellement avec la profondeur de l'arbre, rendant la compréhension des règles de plus en plus difficile. Un arbre d'une profondeur de 1 comprend deux feuilles, mais cette quantité double avec chaque ajout de niveau, rendant la tâche de suivre les règles de décision de plus en plus compliquée à mesure que l'arbre grandit.

Enfin, un autre point faible est le surajustement des arbres décisionnels. Ce phénomène apparaît lorsque l'arbre apprend trop bien les données et capture des détails très spécifiques le rendant très large et complexe ainsi que moins performant aux nouvelles données. (Georges, 2023)

Pour contrer ce surajustement (ou overfitting) il y'a ce qu'on appelle le « pruning » qui consiste a stopper l'arbre de grandir afin d'éviter une trop grande complexité, il s'agit alors là d'un "pre-pruning". Ou bien on effectue un "post pruning" qui laisse l'arbre se développer jusqu'à la fin pour ensuite venir supprimer les branches que l'on considère moins importantes. (George, 2023)

2.3.3 Random Forest

Un autre algorithme d'apprentissage automatique largement utilisé, appelé Random Forest Classifier, lancé par Leo Breiman et Adele Cutler, combine les résultats de plusieurs arbres de décision pour produire un seul résultat. Sa facilité d'utilisation ainsi que sa flexibilité à améliorer son adoption permet de résoudre à la fois les problèmes structurels et de régression (IBM, s.d).

De plus, une étude réalisée par Zaki, Khodadadi, Lim et Towfek (2024), enquêtent sur l'utilisation de l'analyse prédictive et de l'apprentissage automatique pour prédire les abonnements aux dépôts immobiliers. Cette étude comprend plusieurs étapes, dont la collecte et l'analyse de données, l'analyse des données analytiques (EDA), l'ingénierie des fonctionnalités, le traitement des données et la mise en

œuvre de techniques d'apprentissage automatique. Sur base d'un ensemble de données de Kaggle, les auteurs appliquent une série de techniques, dont le Random Forest Classifier ainsi que l'arbre déscionel. Zaki et al. (2024), explique ensuite que les algorithmes sont par la suite comparés entre eux selon leurs

- **Positive Predictive Value (PPV)** : représente le rapport entre la proportion de prédictions positives correctes par rapport au total des prédictions positives. Elle permet de déterminer la précision du modèle à pouvoir identifier correctement les cas de souscriptions aux dépôts à terme bancaires. En d'autres termes plus la PPV est élevées, plus les prédictions de résultats positifs sont fiables.

- **Negative Predictive Value (NPV)** : contrairement à la PPV, la NPV illustre la précision du modèle à détecter les instances où il est peu probable qu'il y ait des souscriptions aux dépôts à terme bancaires. Plus la NPV est élevée, plus les prédictions de résultats négatifs sont fiables.

- **F1- Score :** offre une évaluation équilibrée de la performance totale du modèle. Le score F1 est la moyenne harmonique de la précision, (la proportion de prédictions positives qui sont réellement correctes) et le rappel (la proportion de positifs réels qui ont été correctement identifiés par le modèle). Si le F1-score est élevé, cela signifie que la prédiction ainsi que le rappel le sont aussi, le modèle fonctionne bien tant en précision qu'en rappel.

- **Accuracy :** évalue la justesse des prédictions du modèle. Elle prend en compte à la fois les prévisions optimistes et pessimistes correctes par rapport au total des prédictions.

Tableau 2 : Résultats de classification des différents modèles d'apprentissage automatique

Table 1: Classification result.

Models	P-value PPV	P-value NPV	F1-Score	Accuracy
SGDClassifier	0.868421	0.540984	0.177419	0.44

DOI: https://doi.org/10.54216/AJBOR.110110
Received: July 28, 2023 Revised: October 14, 2023 Accepted: December 12, 2023

KNeighborsClassifier	0.844560	0.903047	0.285714	0.825
LogisticRegression	0.852041	0.917582	0.750000	0.85
GaussianNB	0.896552	0.912281	0.538462	0.85
DecisionTreeClassifier	0.892655	0.915942	0.565217	0.855
RandomForestClassifier	0.878307	0.929972	0.818182	0.875

Source: Zaki, A., Khodadadi, N., Lim, W., & Towfek, S. (2024). Predictive Analytics and Machine Learning in Direct Marketing for Anticipating Bank Term Deposit Subscriptions. American Journal of Business and Operations Research. https://doi.org/10.54216/ajbor.110110

Nous pouvons voir sur le tableau 2 que le Random Forest Classifier possède la meilleure performance avec une précision de 87,5 %, une valeur prédictive négative de 92,9972 % et une valeur prédictive positive de

87,8307 %. Ces résultats permettent aux banques d'avoir des informations précieuses qui leur permettent d'améliorer leurs stratégies marketing dans un environnement financier difficile.

2.3.4 Performance de l'analyse prédictive :

Zhao (2023), discute de l'importance de l'interprétabilité des modèles d'apprentissage automatique dans le contexte bancaire, soulignant que pour que ces modèles soient véritablement utiles aux banques, ils doivent non seulement être précis, mais aussi compréhensibles pour les décideurs. L'interprétabilité est essentielle pour la prise de décisions éclairée et stratégique, permettant aux gestionnaires de comprendre pourquoi un modèle fait certaines prédictions. Cela est crucial dans le secteur bancaire où les décisions peuvent avoir d'importantes répercussions financières et où la confiance dans les modèles utilisés est primordiale. L'article mentionne des méthodes pour améliorer l'interprétabilité, comme l'utilisation de modèles avec une forte interprétabilité intrinsèque ou l'application de techniques d'interprétation modèle-indépendantes après la construction du modèle. Ces approches aident à clarifier les prédictions des modèles d'apprentissage automatique, rendant les insights qu'ils fournissent plus accessibles et applicables à la stratégie marketing des banques.

Dans son article Zhao (2023), effectue une étude dans laquelle le but est de présenter un modèle d'apprentissage automatique interprétable en utilisant les données clients comme exemple, dans le but de prédire si un client bancaire souscrira ou non à un dépôt à terme, offrant ainsi au personnel bancaire une décision efficace pour mener des actions marketing et augmenter les revenus de la banque. Trois aspects principaux de l'analyse interprétable sont abordés dans cet article : Zhao (2023), commence par décrire sa méthodologie de recherche. Il utilise une méthode d'échantillonnage stratifié qui consiste à diviser de manière équilibrer les échantillons positif et négatif entre le training set, le test set ainsi que le vérification set, pour éviter les déséquilibres dans les ensembles de données et assurer une répartition équitable des échantillons positifs et négatifs. Dans cette étude, les échantillons positifs sont les clients qui souscrivent effectivement à un dépôt à terme, tandis que les échantillons négatifs représentent les clients qui ne souscrivent pas à un dépôt à terme.

2.3.4.1 Validation croisée

Pour déterminer la performance du modèle, l'étude utilise la méthode de validation croisée K-fold. Selon Zhao (comme cité dans Zhao, 2023), la méthode de validation croisée K-fold est utilisée pour « évaluer la performance prédictive du modèle d'apprentissage en ensemble lorsqu'il est entraîné. La validation croisée est une méthode de validation de modèle qui évalue l'adaptabilité des algorithmes d'apprentissage automatique aux ensembles de tests, vérifie si le modèle est fiable et prévient le surajustement du modèle. »

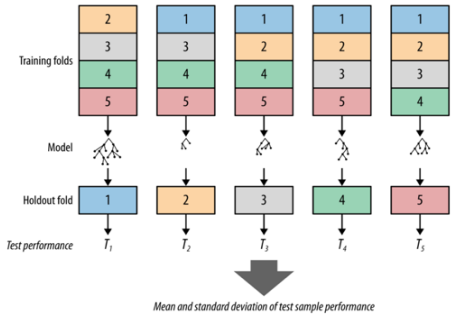
Initialement, l'ensemble des données est séparé en deux parties : un ensemble d'entraînement (D) et un ensemble de test (S). L'ensemble de test (S) est réservé pour l'évaluation finale du modèle. L'ensemble d'entraînement (D) est ensuite divisé en cinq parties égales, ou "plis", nommés D1, D2, D3, D4, et D5. Cette division est essentielle pour les étapes suivantes, où chaque pli servira tour à tour de set de validation.

À chaque itération, un des cinq plis est utilisé comme ensemble de validation et les quatre autres plis sont combinés pour servir d'ensemble d'entraînement. Ainsi, chaque pli est utilisé une fois comme ensemble de validation. Le modèle est entraîné sur les quatre cinquièmes des données et validé sur le cinquième restant. Ce processus est répété cinq fois, une fois pour chaque pli.

Après chaque cycle d'entraînement et de validation, la performance du modèle (par exemple, sa précision) est enregistrée. À la fin des cinq cycles, vous disposez de cinq mesures de performance, une pour chaque fois qu'un pli a servi de set de validation.

Pour obtenir une estimation finale de la performance du modèle, on calcule la moyenne des résultats des cinq expériences. Cette moyenne donne une indication plus fiable de la manière dont le modèle est susceptible de se comporter sur de nouvelles données inédites, comparée à l'utilisation d'un seul ensemble de division entre entraînement et validation.

Figure 6 : Représentation de validation croisée



Source: George, M. (2023). Data Analytics in Marketing - Session 7: Performance & Evaluation. ICHEC.

L'auteur indique que l'ensemble des données utilisées dans l'étude présentait un déséquilibre significatif, avec une faible proportion d'échantillons positifs (11,23 %) par rapport aux échantillons négatifs (88,77 %).

2.3.4.2 Matrice de confusion

La méthode utilisée comme indexe d'évaluation des prédictions du modèle est celle de la matrice de confusion. En effet, d'après Li, Zhao, Shancheng, et Wen (2023), la régression logistique est efficace pour la classification de textes (cité par Zhao, 2023) cette méthode permet est de mieux refléter l'exactitude d'un modèle lorsque celui-ci possède un déséquilibre important entre les échantillons positifs et négatifs.

Tableau 3 : Représentation Matrice de Confusion

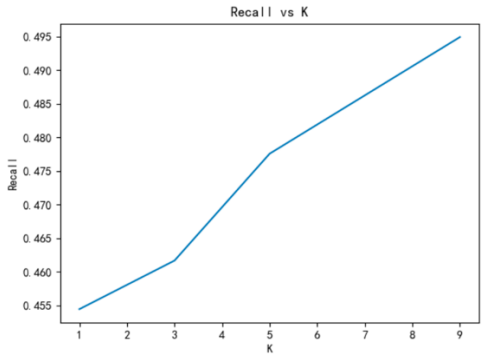
		Actual Class(Observation)	
		Y	**N**
Predicted Class (Expectation)	Y	**TP** (true positive) Correct result	**FP** (false positive) Unexpected result
	N	**FN** (false negative) Missing result	**TN** (true negative) Correct absence of result

Source: George, M. (2023). Data Analytics in Marketing - Session 7: Performance & Evaluation. ICHEC

On peut envisager quatre issues différentes lors de la détermination de l'exactitude avec laquelle un exemple spécifique a été classé: si la classe prédite est Y et la classe réelle est également Y, cela est considéré comme un "Vrai Positif" ou VP; si la classe prédite est Y et la classe réelle est N, cela est un "Faux Positif" ou FP; si la classe prédite est N et la classe réelle est Y, cela est un "Faux Négatif" ou FN; si la classe prédite est N et la classe réelle est également N, cela est un "Vrai Négatif" ou VN (George, 2023).

Ensuite Zhao (2023), utilise l'algorithme des K plus proches voisins (KNN) voir supra. Le choix effectué pour la valeur de K est fait en calculant le score F1 pour les différentes valeurs que K prend. Ainsi la valeur optimale retenue pour K pendant le processus était de 9 comme on peut le voir sur la figure n°5.

Figure 7 : Variation du taux de rappel en fonction du nombre de voisins K pour le modèle KNN

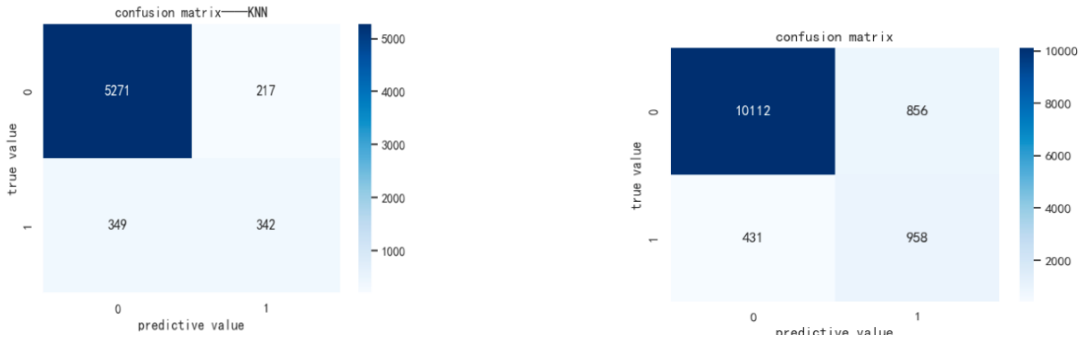


Source: Li, Q., Zhao, S., Zhao, S., & Wen, J. (2023). Logistic regression matching pursuit algorithms for text classification. *Knowledge-Based Systems*

Ensuite, Zhao (2023), utilise également la méthode de régression logistique. La régression logistique est « un type d'analyse de régression utilisé pour prédire la probabilité d'un résultat binaire. Elle modélise la relation entre les variables indépendantes (caractéristiques) et la variable dépendante (cible binaire) en utilisant la fonction logistique. La fonction logistique associe la combinaison linéaire des caractéristiques à une valeur de probabilité comprise entre 0 et 1, représentant la probabilité de la classe positive » (Brightwood, 2024).

Les résultats montrent que la régression logistique a performé mieux que KNN dans cet exemple, car elle a un plus grand nombre de vrais positifs et vrais négatifs et moins de faux négatifs et faux positifs comme on peut le voir dans les deux figures :

Figure 8 : Matrice de Confusion KNN et régression logistique



Source : Zhao, S. (2023). Classification and Prediction of Bank Marketing Activity by Machine Learning. Highlights In Business, Economics And Management, 21, 725-732. https://doi.org/10.54097/hbem.v21i.14752

Dans la matrice de confusion KNN :

- 5271 prédictions sont des vrais négatifs (TN) : le modèle a correctement prédit que ces clients ne s'abonneraient pas.
- 342 prédictions sont des vrais positifs (TP) : le modèle a correctement prédit que ces clients s'abonneraient.
- 217 prédictions sont des faux positifs (FP) : le modèle a incorrectement prédit que ces clients s'abonneraient alors qu'ils ne le feraient pas.
- 349 prédictions sont des faux négatifs (FN) : le modèle a incorrectement prédit que ces clients ne s'abonneraient pas alors qu'ils le feraient.

Dans la matrice de confusion de la régression logistique :

- 10112 prédictions sont des vrais négatifs (TN).
- 958 prédictions sont des vrais positifs (TP).
- 856 prédictions sont des faux positifs (FP).
- 431 prédictions sont des faux négatifs (FN)

2.3.5 Techniques d'amélioration de l'analyse prédictive

2.3.5.1 SMOTE

Dans de nombreux cas, les données peuvent être déséquilibrées, ce qui pose beaucoup de difficultés pour les problèmes de classification. Heureusement, il existe des solutions qui permettent de rééquilibrer ces données. Cette approche est appelée "data-level solutions" et comporte deux formes principales (Côme, 2022).

Dans un premier temps, nous avons le sous-échantillonnage (undersampling). Cette méthode consiste à retirer des individus majoritaires afin que les individus minoritaires ne perdent pas de l'importance. En faisant cela, on réduit la répétition des informations apportées par le grand nombre d'individus (Côme, 2022).

Dans un deuxième temps, nous avons le suréchantillonnage (oversampling). Il s'agit de faire le contraire du sous-échantillonnage. En effet, ici, c'est le nombre d'individus minoritaires qui est augmenté pour qu'ils aient plus d'importance lors de la modélisation. Une technique assez répandue pour cela est la technique du SMOTE (Côme, 2022).

Cette technique est apparue pour la première fois dans un article intitulé "SMOTE : Synthetic Minority Over-sampling Technique" en 2002 dans la revue Journal of Artificial Intelligence Research, écrit par Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall et W. Philip Kegelmeyer. Cette méthode est une technique de suréchantillonnage des observations minoritaires. SMOTE crée de nouveaux individus qui ressemblent aux autres sans pour autant être exactement similaires. Cela diffère de la méthode de clonage où l'on régénère des individus issus de la minorité, mais qui sont identiques (Côme, 2022).

2.3.5.2 Hyperparamètres

Une autre manière d'améliorer un modèle de manière significative est l'optimisation des hyperparamètres. Les hyperparamètres jouent un rôle crucial dans les performances de ces modèles. En attribuant des valeurs adéquates à ceux-ci, on peut fortement améliorer les performances du modèle.

Les hyperparamètres sont des variables externes configurées par les data scientists pour diriger l'entraînement des modèles de machine learning. Parfois appelés hyperparamètres de modèle, ils sont définis manuellement avant de former le modèle. Contrairement aux paramètres, qui sont ajustés automatiquement au cours de l'apprentissage, les hyperparamètres ne sont pas fixés par le processus d'apprentissage (Great Learning Team, 2024).

Des exemples d'hyperparamètres incluent le nombre de nœuds et de couches dans un réseau neuronal ou le nombre de branches dans un arbre de décision. Ces hyperparamètres influencent des aspects cruciaux comme l'architecture du modèle, le taux d'apprentissage et la complexité globale du modèle. (Amazon Web Services, s.d).

Une technique d'ajustement de ces hyperparametres appelé GridSearchCV est utilisée afin de déterminer les valeurs optimales pour un modèle spécifique. Il est important de savoir qu'il est impossible de connaître à l'avance les meilleures combinaisons des hyperparametres et effectuer cela manuellement est un travail beaucoup trop fastidieux d'où l'utilité de GridSearchCV pour automatiser ce processus (Great Learning Team, 2024).

2.4 Data Visualisation

La représentation graphique des données transforme des chiffres et des informations en illustrations visuelles. Cette méthode englobe toutes les stratégies et techniques nécessaires pour présenter les données ou informations de manière illustrée, rendant l'identification, l'analyse et la réponse aux interrogations plus accessibles. Selon le principe qu'"une image vaut mille mots", le cerveau humain traite et comprend les informations visuelles telles que les images et les représentations graphiques plus aisément que les données numériques pures, car il est naturellement plus apte à analyser des visuels. En convertissant les données en formats graphiques, il devient plus simple d'interpréter les résultats et de prendre des décisions éclairées, en exploitant les diverses variables des données pour illustrer différents impacts (Kaur et Geetha, 2022).

Gupta et Singhal (2019), définissent la visualisation des données comme suit : “La visualisation de données est une méthode qui vise à convertir les données efficacement et clairement pour l'utilisateur grâce à une représentation graphique”. La visualisation des données transforme l'exploration de vastes ensembles de données en une expérience visuelle engageante, permettant aux utilisateurs de découvrir des perspectives variées et riches en insights.

Skender et Manevska (2022), ajoutent que « la visualisation elle-même est basée sur la perception rapide des formes visuelles par l'homme, car le cerveau d'une personne moyenne mémorise rapidement les représentations visuelles et les données qu'il recevra de l'environnement à travers le sens visuel. » Elle joue un rôle clé pour permettre de comprendre et d'interpréter de grande quantité de données à l’œil nu, cette technique permet également de faciliter la reconnaissance d'anomalies. Avec l’utilisation de visuel clair, il est possible d’analyser de manière intuitive les données et déceler les tendances de manière plus naturelle. Cet outil puissant ne se contente pas de mettre en lumière les schémas cachés, mais transforme également ces découvertes en connaissances actionnables, renforçant ainsi la prise de décisions. La visualisation permet de simplifier le processus menant de l’analyse à l’action ce qui réduit considérablement les efforts et le temps nécessaire pour pouvoir interpréter et réagir aux informations recueillies (Gupta et Singhal, 2019).

Skender (2023), nous donne des exemples de différent type de visuel que l’on peut retrouver :

- **La matrice de corrélation** : permets d'illustrer clairement les liaisons entre diverses variables en rassemblant d'importantes quantités de données. Dans le cadre d'un ensemble de données.

- **Diagramme circulaire** : illustre les proportions relatives des différentes catégories au sein d'un ensemble de données total, où chaque segment du cercle représente une catégorie de cet ensemble.

- **Les histogrammes** : propose un moyen direct de représenter des informations relatives à d'importants ensembles de données, en classant ces dernières selon des intervalles et en exposant la fréquence d'apparition des valeurs à chaque intervalle, où chaque segment est visualisé sous forme de colonne dans le diagramme.

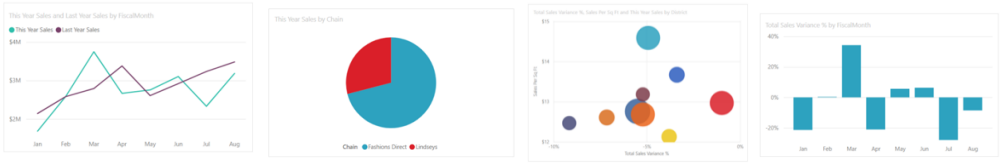
- **Matrice de dispersion** : offrent une méthode de visualisation avancée, révélant les relations multivariantes entre différentes combinaisons de variables au moyen d'un ensemble de graphiques épars, permettant l'ajustement de ces graphiques avec des ellipses de densité pour l'ensemble des données ou spécifiquement pour des groupes de données, en vue d'une analyse approfondie.

- **Les diagrammes de Gantt** : sont utilisés pour organiser et présenter des suites d'activités sur une période déterminée.

- **Les graphiques linéaires** sont principalement utilisés pour afficher les valeurs ou les tendances liées au temps. Les diagrammes de Sankey sont employés pour montrer les flux ou les connexions entre différents segments ou éléments graphiques.

Voici des exemples de visuels représentant un graphique linéaire, un histogramme, une matrice de dispersion ainsi qu’un diagramme circulaire.

Figure 9 : Représentation de différent visuel possible sur PowerBI

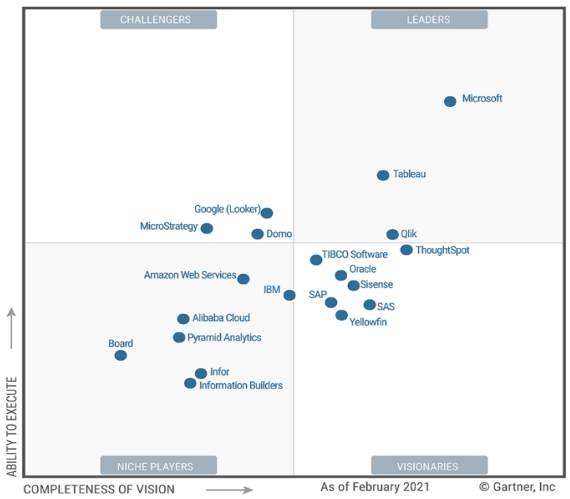


Source: Mihart. (2023, 27 octobre). *Visualization types in Power BI - Power BI*. Microsoft Learn. https://learn.microsoft.com/en-us/power-bi/visuals/power-bi-visualization-types-for-reports-and-q-and-a

De plus, Lennerholt, Van Laere et Söderström (2018), décrit les systèmes BI comme des systèmes d'aide à la décision, offrant un aperçu historique, actuel et prédictif des opérations d'affaires. Il explique que les technologies BI comprennent diverses fonctionnalités telles que la génération de rapports, l'analyse en ligne, l'analytique, l'exploration de données, la gestion de la performance commerciale, l'évaluation comparative, l'exploration de textes et l'analyse prédictive. Il est noté que les systèmes BI traditionnels sont confrontés à des défis liés à l'augmentation du volume des données et à la fréquence d'utilisation, ce qui a mené à l'apparition de l'Intelligence d'Affaires en Libre-Service appelé sous le nom de Self-Service Business Intelligence en anglais (SSBI).
Le SSBI a été introduit par Claudia Imhoff et Colin White (2011), ils affirment avec Lennerholt et al. (2018) que le SSBI est valorisé pour son potentiel à rendre les utilisateurs BI plus autonomes par rapport aux départements informatiques, soulignant l'accès facile, l'analyse et le partage des données avec une moindre dépendance informatique.

Sur le marché des outils de visualisation des données, l'avantage est donné à ceux qui sont les plus intuitifs et les plus proches des besoins des utilisateurs. C'est ce qu'expliquent Skender et Manevska (2022), ils affirment également que la capacité à faire des choix éclairés est cruciale surtout dans un monde où la rapidité devient primordiale. C'est pourquoi la question sur le choix de l'utilisation des outils de visualisation devient évidente. Ainsi on peut voir un afflux important sur le marché des logiciels d'outils de visualisation et d'analyse de données qui favorisent la simplicité et l'accessibilité (Skender et Manevska, 2022).

Figure 10 : Analyse des leaders et challengers dans le domaine des outils de visualisation de données et d'intelligence d'affaires en 2021



Source : Esmoz. (2021, 17 juin). *Tableau VS Power BI : notre comparaison*. https://www.linkedin.com/pulse/tableau-vs-power-bi-notre-comparaison-esmoz-1f/

Cette matrice représentée par la figure n°11 est une matrice appelée “Quadrant Magique de Gartner”, qui permet de rechercher et représenter graphiquement la progression et la position des performances des entreprises ainsi que celle de leurs concurrents dans un marché spécifique basé sur la technologie (Garner, s.d). Elle se comporte de 4 quadrants définis comme suit :

- « Les leaders exécutent bien leur vision actuelle et sont bien placés pour demain » (Garner, s.d).
- « Les visionnaires comprennent où va le marché ou ont une vision pour changer les règles du marché, mais ont du mal à exécuter » (Garner, s.d).
- « Les acteurs de niche se concentrent avec succès sur un petit segment, ou ne sont pas concentrés et n'innovent pas ou ne dépassent pas les autres » (Garner, s.d).
- « Les challengers fonctionnent bien aujourd'hui ou peuvent dominer un large segment, mais ne démontrent pas une compréhension de l'orientation du marché » (Garner, s.d).

Dans ce Quadrant Magique particulier, des entreprises telles que Microsoft et Tableau sont présentées comme des leaders, ce qui indique qu'elles performent bien et ont une vision et une exécution complètes pour leurs outils de visualisation de données. D'autres entreprises, comme Qlik et Salesforce, sont listées comme visionnaires, indiquant une forte innovation, mais potentiellement moins de capacité d'exécution.

Tableau propose une solution de Business Intelligence (BI) particulièrement robuste qui enrichit la visualisation et l'exploration des données pour une variété d'organisations et de profils d'utilisateurs. Avec une interface intuitive qui utilise le glisser-déposer, Tableau permet à ses utilisateurs d'examiner les données cruciales avec aisance, de diffuser des insights pertinents à travers leur entreprise et d'élaborer des visualisations ainsi que des rapports créatifs. De plus, Tableau facilite l'intégration de ses dashboards au sein d'applications commerciales préexistantes, comme SharePoint, ce qui optimise l'analyse rapide. En 2017, Tableau était utilisé par 350 000 personnes (Esmoz, 2021).

De son côté, Microsoft Power BI est une plateforme de cloud Azure spécialisée en analyses et en business intelligence qui centralise et clarifie les données essentielles d'une entité. Elle rend l'analyse et le partage des données plus accessibles pour les utilisateurs en se synchronisant avec diverses sources de données et en fournissant des dashboards modulables. Ces derniers offrent aux utilisateurs la liberté de sélectionner de multiples visualisations et de glisser les données voulues dans le modèle prévu à cet effet. Power BI avait atteint le chiffre de 5 millions d'utilisateurs en 2016 (Esmoz, 2021).

2.4.1 Langues de programmation pour la Data Visualisation

2.4.1.1 Python

Python est l'un des langages les plus populaires en termes d'analyse de données. Il offre un ensemble de bibliothèques très intéressantes comme Matplotlib et Seaborn qui sont conçues spécialement pour la visualisation des données. Matplotlib est la première bibliothèque de visualisation donnée qui prend en charge des graphiques 2D ainsi qu'une prise en charge illimitée des graphiques 3D. Elle permet de créer des figures dans des environnements interactifs sur différentes plateformes et prend également en charge l'animation (Gupta, 2019)

Matplotlib a été créée par John Hunter et de nombreux contributeurs. Devenu aujourd'hui un outil universel et utilisé par de nombreux scientifiques et philosophes, Matplotb s'intègre parfaitement dans le cadre de science des données Python (Sial, Rashdi et Jkan, 2021).

2.4.1.2 Langage R :

R est un langage de programmation conçu spécifiquement pour la visualisation de données. Il offre le support de quatre principaux systèmes de graphiques : basique, treillis, grille, et ggplot2, qui simplifient la représentation de vastes ensembles de données. Les graphiques basiques, en particulier, sont intuitifs et viennent avec une gamme d'outils pratiques pour l'élaboration de présentations descriptives. Utiliser R permet la création efficace de graphiques complexes, économisant ainsi un temps précieux. De plus, ce langage est particulièrement compétent dans la génération de graphiques 3D avancés (Gupta, 2019).

2.4.1.3 Java :

Grâce à Java et à ses diverses bibliothèques intégrées telles que Java 2D, Java 3D, et Java Advanced Imaging, la visualisation des données devient un processus simplifié. L'accès à ces bibliothèques accélère de manière significative le développement des applications dédiées à la visualisation de données. Java permet également l'incorporation de graphiques interactifs dans les applications web, offrant la capacité de créer une multitude de graphiques et de modèles combinés. Son langage de définition graphique avancé assure une personnalisation sans limites. De plus, pour ce qui est de l'analyse de modélisation interactive et de l'intégration de données numériques, VisaAD, une ressource clé parmi les bibliothèques Java, est essentielle (Gupta, 2019).

2.5 Introduction à la méthodologie CRISP-DM

Dans le cadre de mon analyse, je vais utiliser la méthodologie CRISP-DM (Cross Industry Standard Process for Data Mining). Hotz (2024), explique que cette approche étant adoptée pour la première fois

en 1999 vise à standardiser les processus de data mining à travers différentes industries. C'est aujourd'hui la méthode la plus couramment utilisée dans les projets de data mining, d'analytique et de science des données. L'application d'une version flexible du CRISP-DM, combinée à des méthodes de gestion de projet agile adaptées aux équipes, s'est avérée produire les résultats les plus efficaces (Hotz, 2024).

Dans cette partie, nous allons suivre la méthodologie CRISP-DM (Cross Industry Standard Process for Data Mining) pour analyser en profondeur le dataset fourni et développer un modèle prédictif capable de déterminer si un client va souscrire à un dépôt à terme. La méthodologie CRISP-DM, reconnue pour sa robustesse et son efficacité, se compose de plusieurs phases essentielles, que nous allons appliquer de manière structurée.

La méthodologie CRISP-DM est structurée en six phases, comme détaillée par Hotz (2024) :

2.5.1 Business Understanding

Chaque projet efficace débute par une compréhension profonde des nécessités du client, une réalité que la méthodologie CRISP-DM embrasse pleinement. Durant cette phase, l'accent est mis sur la saisie des objectifs et des besoins spécifiques du projet. À l'exception d'une tâche, les trois autres tâches de cette phase constituent des activités de gestion de projet essentielles et communes à la plupart des projets :

- **Identification des objectifs d'affaires** : Il est crucial de saisir clairement, du point de vue de l'entreprise, les aspirations réelles du client, comme le suggère le guide CRISP-DM, pour ensuite établir les critères de réussite commerciale.

- **Analyse de la situation** : Il faut déterminer la disponibilité des ressources et les exigences du projet, évaluer les risques et les mesures de contingence, et effectuer une analyse coûts-avantages.

- **Définition des objectifs de data mining** : Outre les objectifs commerciaux, il est important de définir ce que représente le succès du projet du point de vue technique de l'exploration de données.

- **Élaboration du plan de projet** : Choisir les technologies et les outils appropriés et préparer des plans détaillés pour chaque phase du projet.

Selon Hotz (2024), bien que certaines équipes aient tendance à négliger cette étape cruciale, établir une compréhension approfondie du contexte commercial est fondamental, tout comme le fait de poser les fondations d'une maison.

2.5.2 Data Understanding

La phase suivante est celle de la Compréhension des Données. Prolongeant la compréhension des affaires, cette étape accentue l'identification, la collecte et l'analyse des ensembles de données nécessaires pour réaliser les objectifs du projet. Elle se divise en quatre tâches principales :

- **Collecte des données initiales** : Obtenez les données requises et, si nécessaire, intégrez-les à votre outil d'analyse.

- **Description des données** : Analysez les données et notez leurs caractéristiques apparentes telles que le format, le nombre d'entrées ou l'identification des champs.

- **Exploration des données** : Approfondissez l'étude des données. Interrogez-les, visualisez-les et décelez les relations existantes entre elles.

- **Vérification de la qualité des données** : Évaluez la propreté ou la contamination des données et consignez les éventuels problèmes de qualité.

2.5.3 Data Preparation

D'après Hotz (2024), il est communément admis que la préparation des données représente 80 % du travail d'un projet.

Cette étape, souvent désignée sous le terme de « nettoyage des données », prépare les jeux de données définitifs en vue de leur modélisation. Elle inclut cinq tâches essentielles :

- **Sélection des données** : Choisissez les jeux de données à utiliser et justifiez les motifs de leur sélection ou de leur rejet.

- **Nettoyage des données** : Cette tâche est généralement la plus longue. Elle est cruciale pour éviter les problèmes liés à l'introduction de données erronées, qui pourraient compromettre les résultats. Le processus standard ici consiste à corriger, estimer ou éliminer les valeurs incorrectes.

- **Construction des données** : Générez de nouveaux attributs qui seront bénéfiques pour l'analyse.

- **Intégration des données** : Constituez de nouveaux ensembles de données en fusionnant des informations issues de diverses sources.

- **Formatage des données** : Ajustez le format des données selon les besoins. Vous pourriez, par exemple, transformer des chaînes de caractères contenant des nombres en valeurs numériques pour faciliter les opérations mathématiques.

2.5.4 Modeling

Selon Hortz (2024), cette phase est généralement perçue comme le travail le plus captivant en science des données est aussi souvent la phase la plus brève du projet. Dans cette étape, vous êtes susceptible de développer et d'évaluer une variété de modèles en utilisant différentes techniques de modélisation. Cette phase se divise en quatre tâches essentielles :

- **Choix des techniques de modélisation** : Identifiez les algorithmes à tester, tels que la régression ou les réseaux de neurones.

- **Conception des tests** : Selon votre approche de modélisation, il peut être nécessaire de segmenter vos données en ensembles d'entraînement, de test et de validation.

- **Construction du modèle** : Bien que cela puisse sembler prestigieux, cela peut se résumer à exécuter quelques lignes de code, comme "reg = LinearRegression().fit(X, y)".

- **Évaluation du modèle** : En règle générale, plusieurs modèles sont mis en compétition, et il incombe au data scientist d'interpréter les résultats en fonction de sa connaissance du domaine, des critères de succès établis et du plan de test.

Le guide CRISP-DM recommande de « réitérer la construction et l'évaluation du modèle jusqu'à ce que vous soyez convaincu d'avoir trouvé le ou les meilleurs modèles ». Toutefois, dans la pratique, il est conseillé de poursuivre les itérations jusqu'à obtenir un modèle jugé adéquat, de progresser à travers les différentes étapes du cycle de vie du CRISP-DM, puis de continuer à perfectionner le modèle lors des itérations ultérieures.

2.5.5 Évaluation

Tandis que la tâche d'Évaluation des Modèles de la phase de Modélisation se focalise sur l'analyse technique des modèles, la phase d'Évaluation étend l'analyse pour déterminer quel modèle correspond le mieux aux objectifs commerciaux et définir les actions à suivre. Cette phase se divise en trois tâches principales :

- **Analyse des résultats** : Les modèles satisfont-ils les critères de réussite définis par l'entreprise ? Quel(s) modèle(s) devrait être validé pour utilisation dans l'entreprise ?

- **Examen du processus** : Revoir le travail réalisé. Quelque chose a-t-il été négligé ? Toutes les étapes ont-elles été menées correctement ? Résumer les résultats et apporter les ajustements nécessaires.

- **Planification des étapes suivantes** : Basée sur les trois tâches précédentes, déterminer si l'on doit passer à la phase de déploiement, continuer les itérations ou commencer de nouveaux projets.

2.5.6 Deployement

Selon les besoins, la phase de déploiement peut être aussi simple que la génération d'un rapport ou aussi complexe que la mise en œuvre d'un processus d'exploration de données répétable à travers l'entreprise.

- **Guide CRISP-DM**

Un modèle n'est particulièrement utile que si le client peut accéder à ses résultats. La complexité de cette phase varie considérablement. Cette phase finale comprend quatre tâches :

- **Planifier le déploiement** : Élaborer et documenter un plan pour le déploiement du modèle.

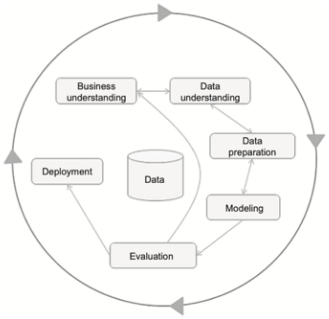
- **Planifier la surveillance et la maintenance** : Développer un plan complet de surveillance et de maintenance pour éviter les problèmes lors de la phase opérationnelle (ou post-projet) du modèle.

- **Produire le rapport final** : L'équipe du projet documente un résumé du projet qui peut inclure une présentation finale des résultats de l'exploration de données.

- **Réviser le projet** : Réaliser une rétrospective du projet pour discuter de ce qui a bien fonctionné, ce qui aurait pu être amélioré et comment s'améliorer à l'avenir.

Hotz (2024), affirme que le travail de votre organisation pourrait ne pas s'arrêter là. En tant que cadre de projet, le CRISP-DM ne détaille pas ce qu'il faut faire après le projet (également connu sous le nom d'« opérations »). Mais si le modèle passe en production, assurez-vous de maintenir le modèle en production. Une surveillance constante et des ajustements occasionnels du modèle sont souvent nécessaires. Selon les exigences spécifiques, la phase de déploiement peut varier d'une simple génération de rapport à l'implémentation d'un processus d'exploration de données répété à l'échelle de l'entreprise.

Figure 11 : Cross Industry Standard Process for Data Mining



Source: Kotu, V., & Deshpande, B. (2019). Data Science: Concepts and Practice (2nd ed.). Morgan Kaufmann. Récupéré de https://asolanki.co.in/wp-content/uploads/2019/04/Data-Science-Concepts-and-Practice-2nd-Edition-3.pd

2.6 Optimisation des Stratégies d'Entreprise par l'Expérimentation Rigoureuse

L'article "The Discipline of Business Experimentation" de Stefan Thomke et Jim Manzi, publié dans le Harvard Business Review, explore l'importance de l'expérimentation scientifique dans le domaine des affaires pour tester et valider des innovations avant de les déployer à grande échelle. Les auteurs soulignent que, contrairement à la science, de nombreuses entreprises ne suivent pas de protocoles rigoureux lorsqu'elles introduisent de nouveaux produits ou services, se fiant souvent à des données historiques qui ne prédisent pas nécessairement les réactions futures des clients (Thomke et Manzi, 2014).

Un exemple marquant est celui de J.C. Penney, où un changement radical de stratégie de vente a conduit à une baisse importante des ventes, car les décisions étaient basées sur des hypothèses non testées. L'article insiste sur l'importance de mener des tests rigoureux, similaires aux essais cliniques qui sont réalisés dans l'industrie pharmaceutique, pour évaluer l'impact de nouvelles stratégies commerciales (Thomke et Manzi, 2014).

Pour mener une expérience efficace, l'article propose plusieurs étapes clés : définir clairement le but de l'expérience, obtenir l'engagement des parties prenantes à respecter les résultats, s'assurer de la faisabilité de l'expérience, et garantir la fiabilité des résultats. Il est crucial de poser des questions précises et de structurer les tests de manière à isoler les variables indépendantes et à observer les effets mesurables (Thomke et Manzi, 2014).

De plus, l'article met en avant l'importance d'utiliser des outils analytiques avancés, y compris les techniques de big data, pour améliorer la validité statistique des expériences, en particulier lorsque les échantillons sont limités. Les auteurs encouragent une approche d'apprentissage continu, où les résultats des expériences sont utilisés pour affiner les hypothèses et guider les décisions stratégiques. Ce processus analytique est également indispensable pour un déploiement progressif, où les ajustements peuvent être faits en fonction des retours du marché (Thomke et Manzi, 2014).

Enfin, l'auteur rajoute que ces étapes sont essentielles pour garantir le succès lors du déploiement d'une nouvelle stratégie ou d'un produit. En effet, le déploiement à grande échelle repose sur la validation d'hypothèses lors de la phase d'expérimentation. Sans cette rigueur, les risques d'échec augmentent considérablement, comme dans le cas de J.C. Penney. De ce fait, les tests rigoureux permettent d'identifier des problèmes potentiels en amont, facilitant ainsi un déploiement plus fluide et efficace.

2.7 Évaluation et Développement de la Maturité Analytique avec le Modèle DELTA Plus

L'article "DELTA Plus Model & Five Stages of Analytics Maturity" du International Institute for Analytics présente un cadre pour évaluer et développer la maturité analytique des entreprises. Ce cadre, connu sous le nom de modèle DELTA Plus, a été élaboré pour aider les organisations à comprendre où elles se situent dans leur parcours analytique et comment elles peuvent progresser pour devenir des concurrentes analytiques de premier plan (Davenport, 2018).

Le modèle DELTA original comprenait cinq éléments essentiels : Données (Data), Orientation d'Entreprise (Enterprise), Leadership, Cibles (Targets), et Analystes. À ces éléments ont été ajoutés deux nouveaux composants, la Technologie et les Techniques Analytiques, pour former le modèle DELTA Plus. Ce modèle met l'accent sur la nécessité d'une intégration et d'une gestion appropriées des données, d'un leadership fort en matière d'analyse, de cibles stratégiques précises, et de compétences analytiques avancées pour réussir dans l'utilisation des données (Davenport, 2018).

Le modèle DELTA Plus est une extension du modèle original DELTA, comprenant sept éléments clés qui déterminent la maturité analytique d'une organisation. Ces éléments sont :

- **Data** : Qualité, accessibilité et intégration des données.
- **Entreprise** : Orientation organisationnelle vers l'analytique, incluant des processus intégrés à l'échelle de l'entreprise.
- **Leadership** : Engagement des dirigeants à soutenir les initiatives analytiques.
- **Targets** : Identification d'objectifs stratégiques pour l'utilisation des analyses.

- **Analysts** : Disponibilité et compétence des talents analytiques.

- **Technologie** : Infrastructure technologique pour soutenir les activités analytiques.

- **Analytical Techniques** : Méthodologies et techniques utilisées pour l'analyse des données.

Les cinq stades de maturité analytique sont décrits par Davenport (2018), comme suit :

- **Analytically Impaired** : Les entreprises à ce stade prennent des décisions basées sur des intuitions sans utilisation formelle de l'analyse.

- **Localized Analytics** : Des initiatives analytiques existent, mais elles sont confinées à des silos organisationnels.

- **Analytical Aspirations** : Les entreprises reconnaissent la valeur de l'analyse et cherchent à améliorer leurs capacités, mais n'ont pas encore réalisé de progrès significatifs.

- **Analytical Companies** : Ces entreprises utilisent les données de manière coordonnée à travers l'organisation, mais n'exploitent pas encore pleinement leur potentiel analytique pour une concurrence stratégique.

- **Analytical Competitors** : Les entreprises à ce stade utilisent les analyses de manière stratégique et généralisée pour obtenir un avantage.

Le modèle DELTA Plus joue un rôle crucial dans la phase de déploiement des stratégies analytiques. Pour qu'une entreprise puisse déployer efficacement ses analyses et en tirer un avantage compétitif, elle doit atteindre un niveau avancé de maturité analytique. Ainsi, en progressant dans les différents stades de maturité, l'organisation sera mieux préparée à intégrer l'analytique dans ses décisions à grande échelle, garantissant ainsi une exécution plus précise et plus rentable des stratégies. Un déploiement réussi dépend non seulement de l'expérimentation, mais aussi de la capacité de l'entreprise à exploiter pleinement ses données et ses outils analytiques.

3 Analyse Exploratoire et Modélisation Prédictive des Données

Dans cette partie, je vais analyser en profondeur un dataset fourni et développer un modèle prédictif capable de déterminer si un client va souscrire à un dépôt à terme ou non. Le but de cette analyse est d'identifier les facteurs clés influençant la décision des clients et d'optimiser les stratégies marketing de la banque.

Je vais commencer par une analyse exploratoire des données (EDA) pour comprendre la structure et les caractéristiques principales du dataset. Cette étape nous permettra de découvrir des tendances, des motifs et des relations entre les variables, ainsi que de repérer les éventuelles anomalies ou valeurs manquantes.

Ensuite, je procéderai à la préparation des données, incluant le nettoyage des données, la transformation des variables catégorielles en variables numériques, et la normalisation des variables numériques. Cette étape est cruciale pour garantir que les données sont prêtes pour l'analyse et la modélisation.

Une fois les données préparées, j'utiliserai des algorithmes de machine learning pour développer un modèle prédictif. Ce modèle nous permettra de prévoir la probabilité qu'un client souscrive à un dépôt à terme, en se basant sur les caractéristiques et les comportements observés dans le dataset. J'évaluerai et validerai le modèle en utilisant des métriques de performance telles que la précision, le rappel et la F1-score.

Enfin, j'analyserai les résultats du modèle pour comprendre ses points forts et ses limites, et proposerai des recommandations pour son intégration dans les campagnes marketing de la banque.

3.1 Business Understanding

Dans le cadre de ce mémoire, la problématique centrale porte sur l'utilisation de l'analyse des données pour cibler efficacement les clients et personnaliser les offres dans les campagnes marketing des banques. Plus spécifiquement, l'objectif est de créer un modèle prédictif permettant de déterminer si un client va souscrire à un dépôt à terme ou non. Cette approche vise à optimiser les stratégies marketing et à augmenter le taux de souscription aux produits financiers proposés par les banques.

Pour les banques, il est impératif de comprendre les aspirations réelles de leurs clients afin de proposer des offres personnalisées qui répondent à leurs besoins. Dans le contexte de ce projet, cela signifie comprendre les facteurs qui influencent la décision des clients à souscrire à un dépôt à terme.

De plus, il est impératif de comprendre les aspirations réelles de leurs clients afin de proposer des offres personnalisées qui répondent à leurs besoins. Dans le contexte de ce projet, cela signifie comprendre les facteurs qui influencent la décision des clients à souscrire à un dépôt à terme. En utilisant le dataset fourni, nous allons analyser les comportements passés des clients et identifier les caractéristiques clés qui influencent leurs décisions.

L'analyse de la situation actuelle implique l'évaluation des ressources disponibles, telles que les compétences techniques, les technologies et les données existantes. Il est essentiel de déterminer la disponibilité des données nécessaires pour entraîner le modèle prédictif. Le dataset fourni contient des informations détaillées sur les caractéristiques démographiques des clients, leurs comportements d'achat et leurs interactions avec les services bancaires.

Outre les objectifs commerciaux, il est important de définir les objectifs techniques du projet. Cela inclut la création d'un modèle prédictif capable de déterminer la probabilité qu'un client souscrive à un dépôt à terme en se basant sur les données fournies. Les métriques de performance pour évaluer le succès du modèle incluront l'accuracy, la précision, le rappel, et la F1-score. Ces mesures permettront de s'assurer que le modèle est à la fois efficace et fiable. Le dataset sera nettoyé, préparé et utilisé pour entraîner différents algorithmes de machine learning, tels que les arbres de décision, les forêts aléatoires et les réseaux de neurones.

3.2 Data Understanding :

3.2.1 Collecte des Données Initiales

- Source des Données

Le dataset utilisé pour cette analyse provient de Kaggle, une plateforme bien connue pour ses compétitions de science des données et ses vastes ressources de datasets publics. Les données ont été publiées par une institution bancaire anonyme qui a collecté des informations à partir d'interactions et de transactions avec ses clients dans le cadre de ses campagnes de marketing pour des dépôts à terme.

- Sélection des Données

Les données incluent des informations sur les clients qui ont été contactés par la banque pendant une série de campagnes marketing. La sélection des données pour le dataset semble avoir été guidée par l'intérêt de comprendre les facteurs influençant la décision des clients de souscrire à un dépôt à terme. Le dataset est représentatif d'une variété de clients en termes d'âge, de statut professionnel, d'éducation, et de situation financière.

- Collecte et Compilation

Les données semblent avoir été collectées à travers divers canaux de communication, notamment des appels téléphoniques et des rencontres en personne, comme en témoignent les variables telles que le type de contact et la durée des appels. Après la collecte, les données ont été compilées en un fichier CSV, structuré et prêt à être analysé.

- Intégration des Données

Bien que les données aient déjà été nettoyées et préparées par les auteurs du dataset sur Kaggle, je vais tout de même procéder à une vérification et à un nettoyage supplémentaire de mon côté. Cela permettra de garantir que toutes les éventuelles anomalies résiduelles, erreurs d'entrée ou valeurs

manquantes non détectées sont corrigées. Cette étape est cruciale pour assurer une analyse encore plus fiable et cohérente, en prenant en compte tous les aspects techniques spécifiques à notre analyse actuelle

- Confidentialité et Éthique

Étant donné que le dataset a été publié sur une plateforme publique, il est présumé que toutes les données sensibles ont été anonymisées ou que le consentement a été obtenu conformément aux normes éthiques et légales en vigueur. Cela inclut la suppression ou la modification des identifiants personnels pour protéger la vie privée des clients dont les données sont analysées.

- Préparation pour l'Analyse

Enfin, le dataset a été structuré de manière à faciliter l'analyse des comportements des clients et l'évaluation de l'efficacité des campagnes marketing. Des variables comme la durée des appels, le nombre de contacts, et les résultats des campagnes précédentes ont été soigneusement enregistrées et indexées pour permettre une analyse détaillée et une modélisation prédictive des décisions des clients.

- Description des données

Je vais maintenant détailler les différentes variables présentes dans le dataset et fournir une description approfondie de celui-ci. Cette information est cruciale pour comprendre les dimensions et le contexte des données avec lesquelles nous travaillons, permettant ainsi une analyse plus informée et ciblée.

La description des variables m'aidera à identifier les éléments clés qui peuvent influencer la décision d'un client de souscrire à un dépôt à terme, tandis que la compréhension du dataset dans son ensemble nous permettra de mieux appréhender les défis et opportunités associés à la modélisation prédictive dans le domaine bancaire. Voici un tableau récapitulatif des variables, suivi par une description détaillée de chaque composante et du dataset.

Tableau 4 : Description Détaillée des Variables du Dataset

Variable	Type	Description
Age	Numérique	Âge du client en années.
Job	Catégorielle	Type d'emploi du client (ex : admin., blue-collar, entrepreneur, etc.).
Martial	Catégorielle	Statut marital du client (ex : married, single, divorced).
Education	Catégorielle	Niveau d'éducation du client (ex : basic.4y, high.school, university.degree, etc.).
Default	Catégorielle	Indique si le client a un crédit en défaut (oui, non, inconnu).

Housing	Catégorielle	Indique si le client a un prêt immobilier (oui, non, inconnu).
Loan	Catégorielle	Indique si le client a un prêt personnel (oui, non, inconnu).
Contract	Catégorielle	Type de communication utilisé lors du dernier contact (telephone, cellular).
Month	Catégorielle	Mois du dernier contact avec le client.
Day_of_week	Catégorielle	Jour de la semaine du dernier contact avec le client.
Duration	Numérique	Durée du dernier contact, en secondes. Cette variable est fortement influente et doit être utilisée avec prudence, car elle n'est pas connue avant l'appel.
Campaign	Numérique	Nombre de contacts effectués pendant cette campagne et pour ce client.
Pdays	Numérique	Nombre de jours écoulés depuis le dernier contact du client lors d'une campagne précédente (999 signifie que le client n'a pas été contacté précédemment).
Previous	Numérique	Nombre de contacts réalisés avant cette campagne pour ce client.
Poutcome	Catégorielle	Résultat de la campagne marketing précédente (ex : failure, nonexistent, success).
Empvarrate	Numérique	Taux de variation de l'emploi - indicateur trimestriel (ex : -1.8, 1.1). Cela reflète les conditions économiques actuelles.
Conspriceidx	Numérique	Indice des prix à la consommation - indicateur mensuel, reflète l'inflation perçue.
Consconfidx	Numérique	Indice de confiance des consommateurs - indicateur mensuel, mesure le degré d'optimisme que les consommateurs ressentent sur l'état de l'économie.
Euribor3m	Numérique	Taux Euribor à 3 mois - indicateur quotidien, reflète le coût de l'emprunt.
Nremployed	Numérique	Nombre de personnes employées - indicateur trimestriel reflètent la santé du marché du travail.
Y	Catégorielle	Indique si le client a souscrit à un dépôt à terme (oui, non).

3.2.2 Exploration des données

Pour rappel, l'objectif de cette phase est de détecter des patterns, identifier des anomalies, tester des hypothèses et vérifier des suppositions à l'aide de statistiques descriptives et de visualisations graphiques.

Dans notre contexte, l'analyse exploratoire aidera à comprendre les profils des clients d'une banque qui sont susceptibles de souscrire à un dépôt à terme. Cela comprend l'examen de variables telles que l'âge, le type d'emploi, le statut marital, le niveau d'éducation, ainsi que des informations plus spécifiques sur les interactions des clients avec la banque lors des campagnes de marketing.

Voici les étapes que je vais suivre pour l'analyse :

- **Visualiser les distributions** des variables numériques et catégorielles pour observer leur comportement.

- **Analyser les corrélations** entre les variables numériques, en particulier celles qui sont susceptibles d'influencer la souscription aux dépôts à terme.

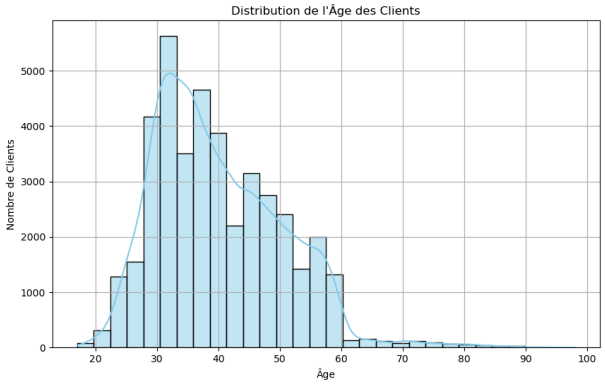
- **Visualiser les relations potentielles** entre les variables clés.

- **Identifier les valeurs manquantes et les valeurs aberrantes** dans le dataset.

3.2.2.1 Distribution Démographique des Clients

Pour débuter, nous explorons la distribution de l'âge des clients. L'âge est un facteur déterminant dans le comportement financier, influençant la propension d'un individu à investir dans des dépôts à terme. Nous nous attendons à voir différentes tendances de souscription en fonction des tranches d'âge.

Figure 12 : Distribution de l'Âge des Clients



Les principales observations du graphique sont les suivantes :

1. **Concentration autour de la Trentaine** : La distribution de l'âge des clients montre une prédominance des individus âgés de 30 à 40 ans. Cette tendance suggère que ce groupe d'âge constitue une part importante de la clientèle bancaire, potentiellement en raison de leur stabilité financière et de leur intérêt pour des produits d'épargne à long terme comme les dépôts à terme.

2. **Pic de Fréquence à 30 Ans** : Le graphique révèle un pic notable autour de l'âge de 30 ans, indiquant que cette tranche d'âge est particulièrement active ou réceptive aux offres bancaires. Cela pourrait être dû à des facteurs tels que l'établissement de carrières, l'achat de logements ou la planification de familles, qui encouragent une gestion financière plus rigoureuse.

3. **Réduction Progressive après 50 Ans** : On observe une diminution progressive du nombre de clients après l'âge de 50 ans. Cela pourrait refléter une moindre demande pour certains produits financiers, ou bien une satisfaction déjà atteinte avec les services bancaires existants, rendant moins nécessaire la souscription à de nouveaux produits.

4. **Faible Représentation des Très Jeunes et des Très Âgés :** Les tranches d'âge en dessous de 20 ans et au-dessus de 60 ans sont moins représentées dans le graphique. Pour les plus jeunes, cela peut s'expliquer par une entrée plus tardive dans le monde financier actif. Pour les plus âgés, cela peut être dû à des revenus fixes ou à une préférence pour des produits financiers différents de ceux analysés ici.

5. **Implications Marketing** : Cette distribution peut guider les banques dans la personnalisation de leurs approches marketing pour différents groupes d'âge. Par exemple, les clients dans la tranche des 30-40 ans pourraient bénéficier de produits d'investissement à long terme ou de prêts hypothécaires. En revanche, les plus jeunes pourraient être intéressés par des comptes d'épargne ou des prêts étudiants, tandis que les plus âgés pourraient préférer des produits de retraite ou des services de gestion de patrimoine.

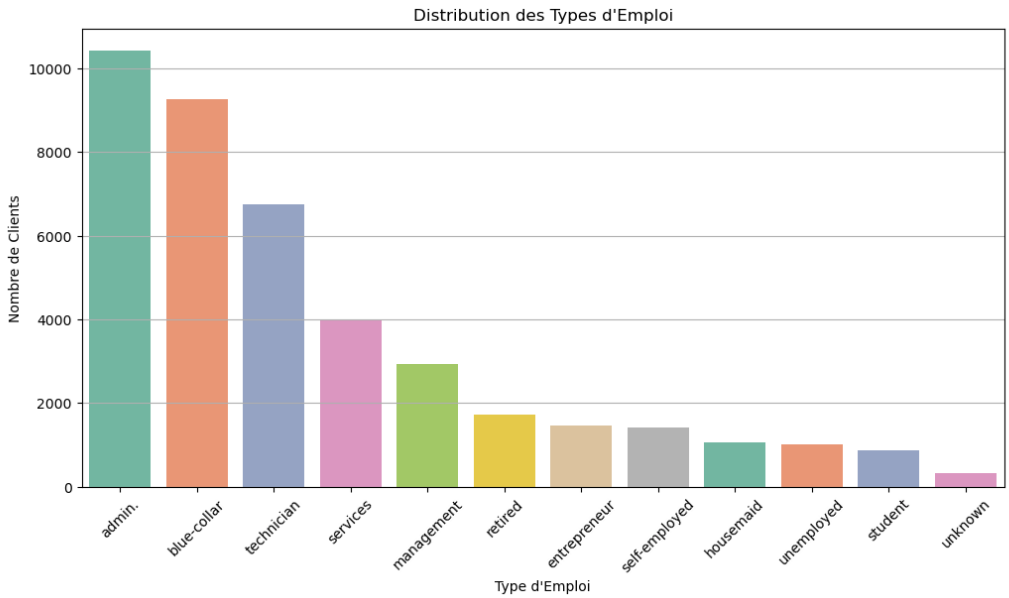
6. **Stratégie et Planification** : Comprendre la répartition par âge permet aux équipes marketing de planifier plus stratégiquement l'allocation des ressources. En concentrant les efforts sur les segments les plus représentés (comme les 30-40 ans), tout en développant de nouvelles stratégies pour atteindre les segments moins représentés, les banques peuvent optimiser leurs campagnes marketing et améliorer leur retour sur investissement.

Conclusion

En résumé, le graphique de la distribution de l'âge des clients offre un aperçu précieux de la composition de la clientèle bancaire. Cette analyse peut guider la banque dans le développement de stratégies marketing plus efficaces et personnalisées, répondant aux besoins variés de ses clients. En se basant sur ces données, les banques peuvent non seulement cibler plus précisément leurs offres, mais aussi anticiper les besoins futurs de leurs clients en fonction des tendances démographiques.

3.2.2.2 Distribution des Types d'Emploi

Figure 13 : Distribution des Types d'Emploi



Les principales observations du graphique sont les suivantes :

1. **Prédominance de Certaines Catégories** : Les catégories d'emploi telles qu'**admin.**, **blue-collar**, **technician**, et **services** sont les plus représentées. Cette tendance suggère que ces groupes professionnels pourraient être les principaux destinataires ou les plus intéressés par les produits bancaires offerts, tels que les dépôts à terme.

2. **Catégories Dominantes** : Les employés administratifs (**admin.**) et les ouvriers (**blue-collar**) constituent les plus grandes proportions du graphique. Cela peut indiquer que les stratégies marketing actuelles sont particulièrement efficaces auprès de ces groupes ou que ces segments de la population sont naturellement plus enclins à investir dans des produits d'épargne à long terme en raison de leurs besoins de sécurité financière et de leur stabilité de revenu.

3. **Moins de Représentation pour Certains Groupes** : Les catégories **retired**, **unemployed**, et **student** montrent des nombres moins élevés, ce qui pourrait refléter un manque d'intérêt ou des obstacles spécifiques à ces groupes, tels que des revenus limités ou des priorités financières différentes qui les rendent moins susceptibles de souscrire à des dépôts à terme.

4. **Implications Marketing** : Cette distribution peut guider les banques dans la personnalisation de leurs approches marketing pour différents segments professionnels. Par exemple, les étudiants et les retraités pourraient bénéficier de messages et d'offres spécifiquement adaptés à leurs situations financières et à leurs besoins potentiellement augmentant leur engagement et leur taux de souscription.

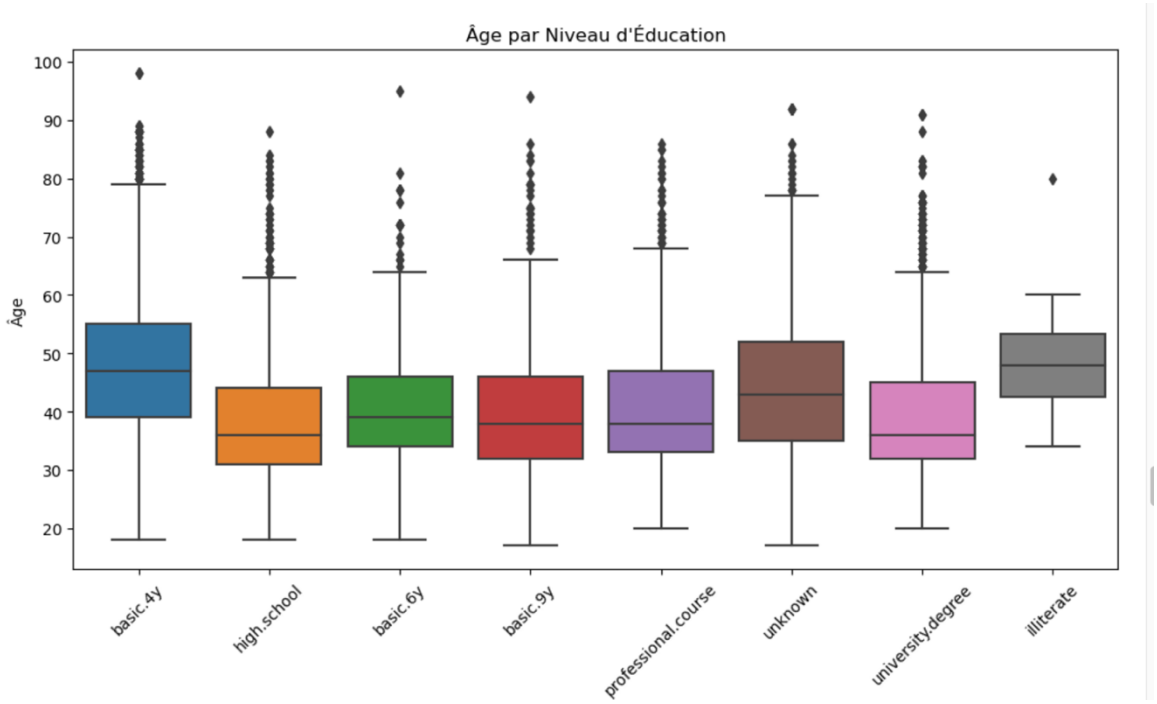
5. **Stratégie et Planification** : Comprendre quelle catégorie professionnelle est plus ou moins engagée permet aux équipes marketing de planifier plus stratégiquement l'allocation des ressources, en concentrant les efforts là où ils sont le plus susceptibles de générer un retour sur investissement positif, tout en développant de nouvelles stratégies pour atteindre les segments sous-représentés.

Conclusion

En résumé, le graphique de la distribution des types d'emploi offre un aperçu précieux de la composition de la clientèle bancaire. Cette analyse peut guider la banque dans le développement de stratégies marketing plus efficaces et personnalisées, tout en répondant aux besoins variés de ses clients.

3.2.2.3 Distribution de l'Âge par Niveau d'Éducation

Figure 14 : Distribution de l'Âge par Niveau d'Éducation



Les principales observations du graphique sont les suivantes :

1. **Diversité des Âges par Niveau d'Éducation** : Le graphique boxplot montre une grande diversité d'âges dans chaque catégorie de niveau d'éducation. Les catégories "basic.4y", "high.school" et "university.degree" ont une répartition relativement large, ce qui indique que ces niveaux d'éducation incluent des individus de différentes tranches d'âge.

2. **Catégories Dominantes** : Les niveaux d'éducation "basic.4y" et "high.school" montrent une concentration plus élevée de clients dans les tranches d'âge intermédiaires, notamment entre 30

et 50 ans. Cela peut indiquer que ces groupes sont plus représentés dans la base de données de la banque et pourraient être des cibles prioritaires pour les campagnes marketing.

3. **Variabilité par Niveau d'Éducation** : La catégorie "university.degree" présente une variation d'âge plus limitée par rapport aux autres catégories, avec une concentration notable autour de 30 à 50 ans. Cette limitation de la variabilité pourrait signifier que les personnes ayant un diplôme universitaire ont des profils d'âge plus uniformes en termes de demande pour des produits financiers.

4. **Présence d'Anomalies** : Certaines catégories, comme "unknown" et "illiterate", montrent une présence d'âges extrêmes ou de valeurs aberrantes. Par exemple, la catégorie "illiterate" a un client dans la tranche des 90 ans, ce qui est une exception notable. Ces anomalies peuvent nécessiter une analyse plus approfondie pour comprendre leur impact sur les stratégies marketing.

5. **Implications Marketing** : La distribution par niveau d'éducation peut guider les banques dans la personnalisation de leurs approches marketing. Par exemple, les clients avec un niveau d'éducation "high.school" ou "university.degree" pourraient être intéressés par des produits d'épargne et d'investissement à long terme, tandis que ceux avec un niveau "basic.4y" pourraient bénéficier de produits plus accessibles et de conseils financiers personnalisés.

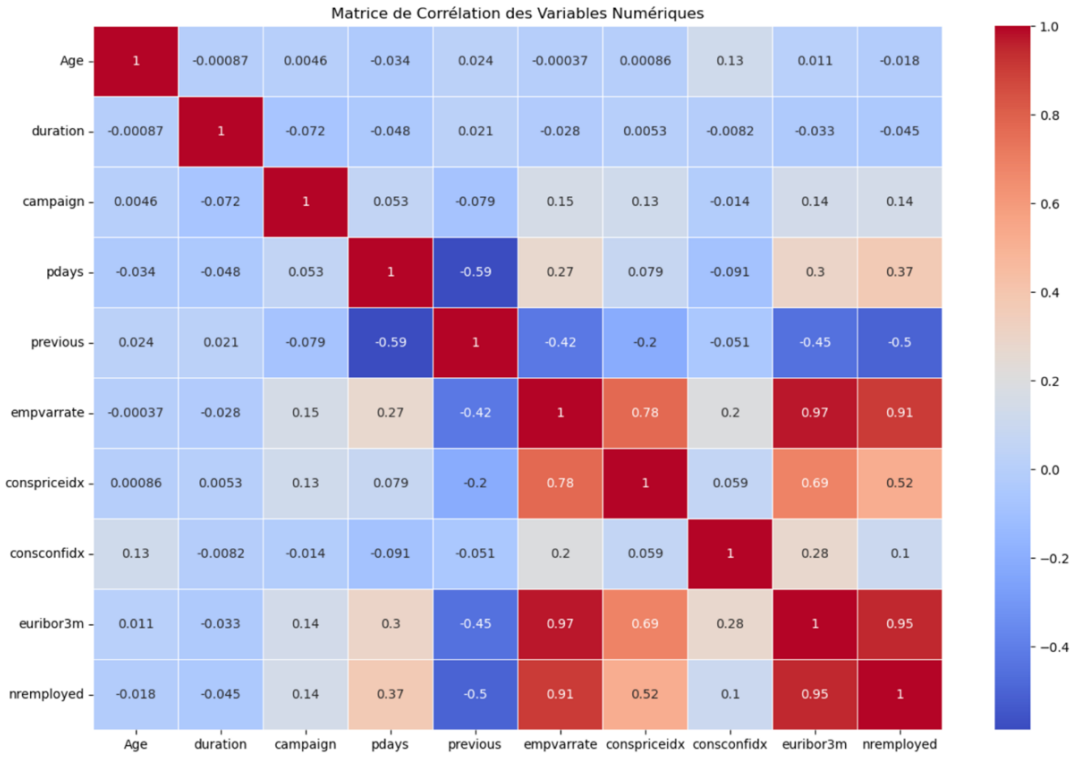
6. **Stratégie et Planification :** Comprendre la répartition par âge et niveau d'éducation permet aux équipes marketing de planifier plus stratégiquement l'allocation des ressources. En concentrant les efforts sur les segments d'éducation les plus représentés et en développant des stratégies pour atteindre les segments moins représentés, les banques peuvent optimiser leurs campagnes marketing et améliorer leur retour sur investissement.

Conclusion

En résumé, le graphique de la distribution de l'âge par niveau d'éducation offre un aperçu précieux de la composition de la clientèle bancaire en termes de tranche d'âge et de niveau d'éducation. Cette analyse peut guider la banque dans le développement de stratégies marketing plus efficaces et personnalisées, répondant aux besoins variés de ses clients. En se basant sur ces données, les banques peuvent cibler plus précisément leurs offres et anticiper les besoins futurs de leurs clients en fonction des tendances éducatives et démographiques.

3.2.2.4 Matrice de corrélation

Figure 15 : Matrice de Corrélation



Les principales observations du graphique sont les suivantes :

1. **Corrélations Faibles et Fortes** : La matrice de corrélation montre les relations entre différentes variables numériques du dataset. Les valeurs de corrélation vont de -1 à 1, où 1 indique une corrélation positive parfaite, -1 une corrélation négative parfaite, et 0 aucune corrélation. Par exemple, les variables "euribor3m" et "nremployed" montrent une forte corrélation positive (0.91), ce qui suggère que lorsque l'indice Euribor à 3 mois augmente, le nombre d'employés augmente également.

2. **Corrélation Négative Significative** : Une forte corrélation négative est observée entre "previous" et "pdays" (-0.59). Cela indique que plus le nombre de jours écoulés depuis le dernier contact est élevé, moins le client a été contacté précédemment. Cette relation est intuitive, car un contact récent réduit naturellement le nombre de jours depuis le dernier contact.

3. **Relations Faibles** : Certaines variables montrent des relations très faibles entre elles, telles que "Age" et "duration" (-0.0087) ou "duration" et "empvarrate" (-0.028). Ces faibles corrélations suggèrent qu'il n'y a pas de lien direct et significatif entre ces paires de variables.

4. **Interprétation des Variables Économiques** : Les variables économiques telles que "euribor3m", "empvarrate", et "conspriceidx" montrent des corrélations significatives entre elles. Par exemple, "empvarrate" (taux de variation de l'emploi) et "euribor3m" ont une corrélation positive de 0.67, indiquant une relation entre les taux d'intérêt et les taux de variation de l'emploi. Ces relations

peuvent être utiles pour comprendre l'impact des conditions économiques sur le comportement des clients.

5. **Impact sur le Modèle Prédictif** : Comprendre les corrélations entre les variables est crucial pour le développement de modèles prédictifs. Les variables fortement corrélées peuvent parfois être redondantes et conduire à un surajustement du modèle. Par exemple, "euribor3m" et "nremployed" ayant une corrélation de 0.91, l'inclusion des deux variables pourrait nécessiter une attention particulière pour éviter le surajustement.

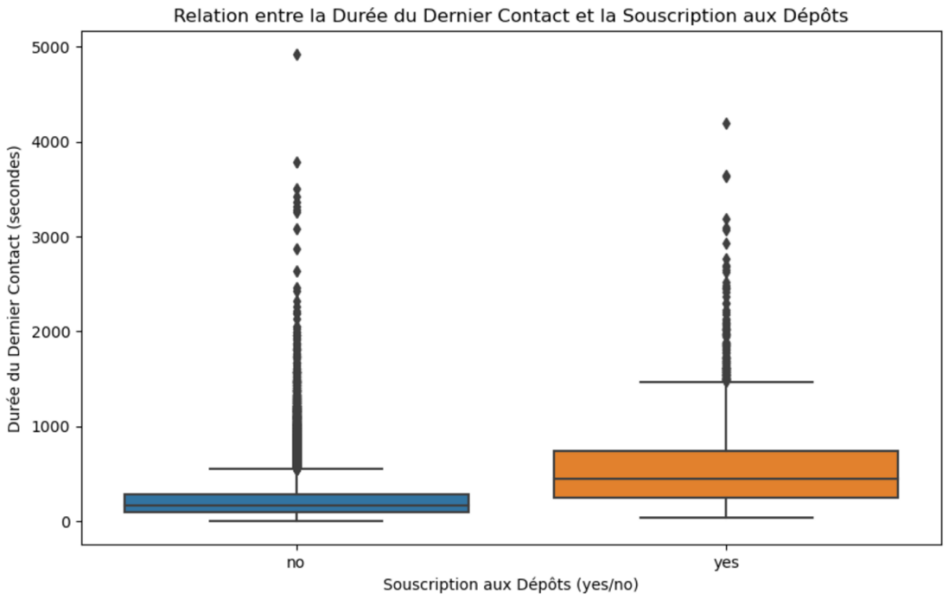
6. **Identification des Relations Clés** : La matrice de corrélation permet d'identifier les variables qui pourraient avoir un impact significatif sur les modèles prédictifs. Par exemple, la durée des appels ("duration") montre une légère corrélation positive avec le succès des campagnes ("previous"), suggérant que des appels plus longs pourraient être associés à une meilleure réponse des clients.

Conclusion

En résumé, la matrice de corrélation des variables numériques offre un aperçu précieux des relations entre les différentes caractéristiques du dataset. Cette analyse peut guider la banque dans la sélection des variables les plus pertinentes pour le développement de modèles prédictifs efficaces et dans la compréhension des dynamiques économiques influençant le comportement des clients. En se basant sur ces corrélations, les banques peuvent optimiser leurs stratégies marketing et améliorer leurs taux de souscription en ciblant les facteurs les plus influents.

3.2.2.5 Analyse de la Relation entre la Durée du Dernier Contact et la Souscription aux Dépôts

Figure 16 : Relation entre la Durée du Dernier Contact et la Souscription aux Dépôts



Les principales observations du graphique sont les suivantes :

1. **Médianes des Durées de Contact** : La médiane de la durée du dernier contact pour les clients ayant souscrit à un dépôt est plus élevée que pour ceux qui n'ont pas souscrit. Cela signifie que, typiquement, les conversations réussies qui aboutissent à une souscription sont plus longues, ce qui pourrait indiquer un besoin de temps supplémentaire pour convaincre le client ou pour fournir des informations détaillées sur le produit.

2. **Présence de Valeurs Extrêmes** : Les deux groupes présentent des valeurs extrêmes (outliers), mais ces valeurs sont plus fréquentes et s'étendent sur une plus grande plage dans le groupe des non-souscriptions. Cela peut indiquer qu'il existe des appels exceptionnellement longs qui n'aboutissent pas à une souscription, possiblement en raison de la réticence des clients ou de discussions prolongées sans succès.

3. **Implications Marketing** : Comprendre que les appels plus longs sont associés à des taux de souscription plus élevés peut guider les équipes marketing dans l'élaboration de scripts d'appel plus complets et détaillés. Les agents peuvent être formés pour engager les clients dans des conversations plus longues et informatives, augmentant ainsi les chances de conversion.

4. **Stratégie d'Engagement** : Les résultats de cette analyse peuvent également influencer la stratégie d'engagement des clients. Par exemple, les campagnes marketing pourraient être ajustées pour allouer plus de temps aux agents pour chaque appel, afin de maximiser les interactions et potentiellement augmenter les taux de souscription.

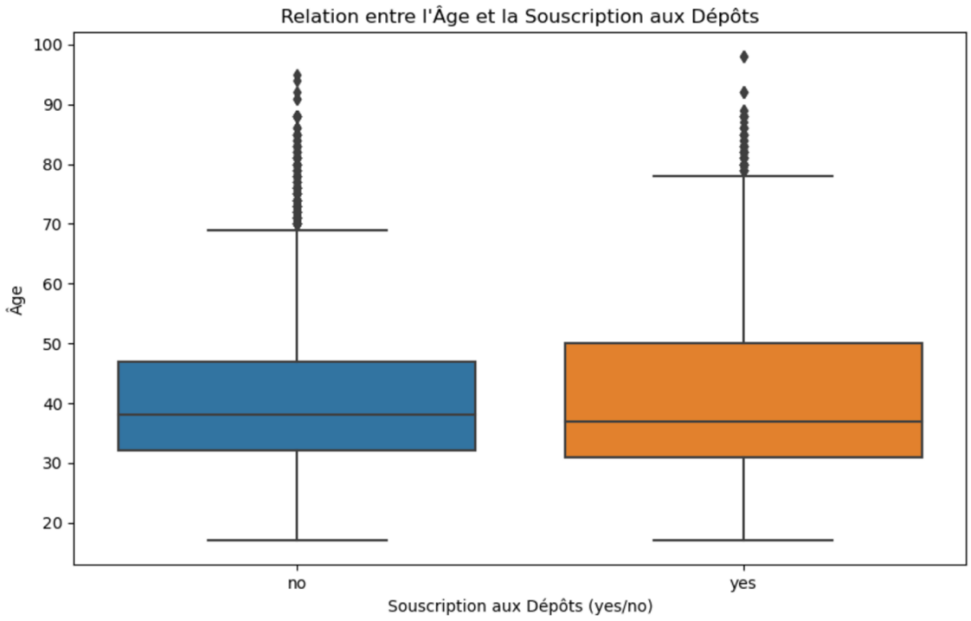
5. **Optimisation des Ressources** : En comprenant que des interactions plus longues tendent à être plus fructueuses, les banques peuvent optimiser l'allocation des ressources humaines en priorisant les appels de qualité sur la quantité. Cela peut également justifier l'investissement dans des formations approfondies pour les agents afin de les rendre plus efficaces dans leurs interactions avec les clients.

Conclusion

En résumé, le graphique de la relation entre la durée du dernier contact et la souscription aux dépôts révèle que les interactions plus longues sont généralement plus efficaces pour convertir les prospects en souscripteurs de dépôts à terme. Cette analyse offre des insights précieux pour les équipes marketing et de vente, leur permettant de développer des stratégies d'engagement plus efficaces et de mieux allouer leurs ressources pour maximiser les taux de souscription.

3.2.2.6 Analyse de la Relation entre l'Âge et la Souscription aux Dépôts

Figure 17 : Relation entre l'Âge et la Souscription aux Dépôts



Les principales observations du graphique sont les suivantes :

1. **Distribution de l'Âge par Souscription :** Le graphique boxplot montre la distribution des âges des clients en fonction de leur souscription à un dépôt à terme. Les deux groupes ("yes" et "no") présentent des répartitions d'âge similaires, avec des médianes proches autour de la quarantaine. Cependant, il y a des différences subtiles qui peuvent être pertinentes pour les stratégies marketing.

2. **Âge Médian Légèrement Plus Élevé pour les Souscriptions :** La médiane de l'âge des clients qui ont souscrit à un dépôt ("yes") est légèrement plus élevée que celle des clients qui n'ont pas souscrit ("no"). Cela peut indiquer que les clients plus âgés sont légèrement plus enclins à souscrire à des dépôts à terme, peut-être en raison de leur intérêt accru pour des options d'épargne et de sécurité financière à long terme.

3. **Variabilité des Âges** : Les deux groupes montrent une grande variabilité d'âge, avec des clients allant de jeunes adultes à des personnes âgées. Cette diversité suggère que la souscription à des dépôts à terme n'est pas limitée à une tranche d'âge spécifique, mais peut être influencée par d'autres facteurs tels que la situation financière et les objectifs d'épargne.

4. **Présence de Valeurs Extrêmes :** Les deux groupes présentent des valeurs extrêmes, avec des clients âgés de plus de 90 ans dans les deux catégories. Ces valeurs extrêmes, bien que rares, montrent que même les clients très âgés peuvent être intéressés par des dépôts à

terme, soulignant l'importance de ne pas négliger aucun segment d'âge dans les campagnes marketing.

5. **Implications Marketing** : Comprendre que la souscription à des dépôts à terme est légèrement plus élevée chez les clients plus âgés peut guider les équipes marketing à cibler ce groupe avec des produits adaptés à leurs besoins spécifiques, tels que des solutions de retraite ou des plans d'épargne sécurisés. Les campagnes peuvent être conçues pour mettre en avant la sécurité et les avantages à long terme des dépôts à terme pour les clients d'âge moyen avancé.

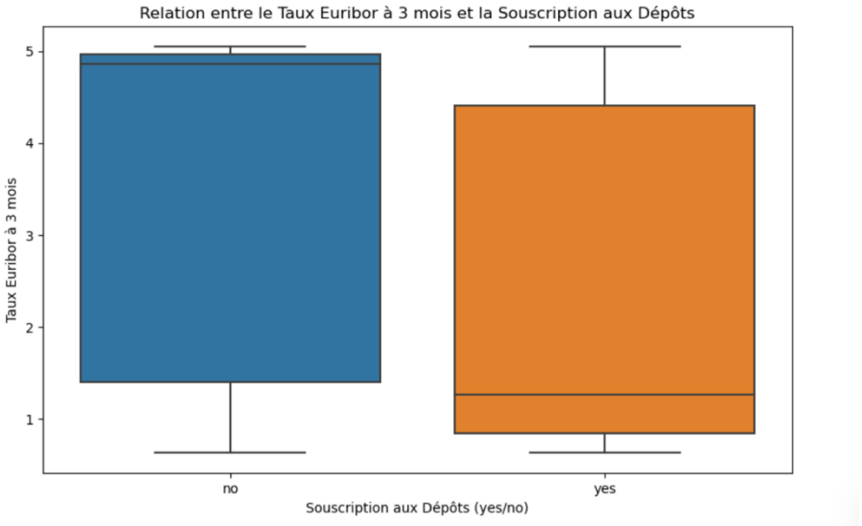
6. **Stratégie de Segment** : La segmentation par âge peut aider les banques à développer des messages marketing plus personnalisés. Par exemple, les jeunes adultes pourraient être attirés par des offres combinées incluant des dépôts à terme et des conseils financiers pour la planification à long terme, tandis que les clients plus âgés pourraient bénéficier d'une approche axée sur la stabilité financière et la sécurité des investissements.

Conclusion

En résumé, le graphique de la relation entre l'âge et la souscription aux dépôts révèle que bien que la souscription soit relativement uniforme à travers les âges, les clients plus âgés montrent une légère tendance à souscrire davantage. Cette analyse offre des insights précieux pour les équipes marketing, leur permettant de développer des stratégies d'engagement plus ciblées et d'optimiser les offres pour répondre aux besoins spécifiques de chaque groupe d'âge. En se basant sur ces données, les banques peuvent améliorer leurs taux de souscription et renforcer leur relation avec les clients à différents stades de la vie.

3.2.2.7 Analyse de la Relation entre le Taux Euribor à 3 mois et la Souscription aux Dépôts

Figure 18 : Relation entre le Taux Euribor à 3 mois et la Souscription aux Dépôts



Les principales observations du graphique sont les suivantes :

1. **Différence de Distribution entre les Groupes** : Le graphique boxplot montre que les clients ayant souscrit à un dépôt à terme ("yes") tendent à le faire lorsque le taux Euribor à 3 mois est plus bas, avec une médiane autour de 1.5. En revanche, les clients n'ayant pas souscrit ("no") montrent une répartition de taux plus élevée, avec une médiane autour de 4.5.

2. **Variation des Taux** : La variation des taux Euribor pour les non-souscripteurs couvre presque toute la plage de 1 à 5, tandis que pour les souscripteurs, la majorité des taux est concentrée entre 1 et 3. Cela suggère que des taux Euribor plus bas peuvent être plus favorables à la souscription de dépôts à terme, probablement en raison de conditions de prêt plus attractives.

3. **Implications des Taux Bas** : Les taux Euribor plus bas semblent être associés à une plus grande probabilité de souscription. Cela peut être dû au fait que des taux plus bas réduisent le coût d'opportunité des dépôts à terme par rapport à d'autres formes d'investissement ou de dépense.

4. **Présence de Valeurs Extrêmes** : Les deux groupes présentent des valeurs extrêmes, mais ces valeurs sont plus fréquentes et s'étendent sur une plus grande plage dans le groupe des non-souscriptions. Cela peut indiquer que certains clients sont influencés par d'autres facteurs que le taux Euribor lorsqu'ils décident de ne pas souscrire.

5. **Implications Marketing** : Comprendre que les souscriptions augmentent avec des taux Euribor plus bas peut guider les banques à adapter leurs offres en fonction des conditions du marché. Par exemple, durant les périodes de taux bas, les banques pourraient intensifier leurs efforts de marketing pour promouvoir les dépôts à terme.

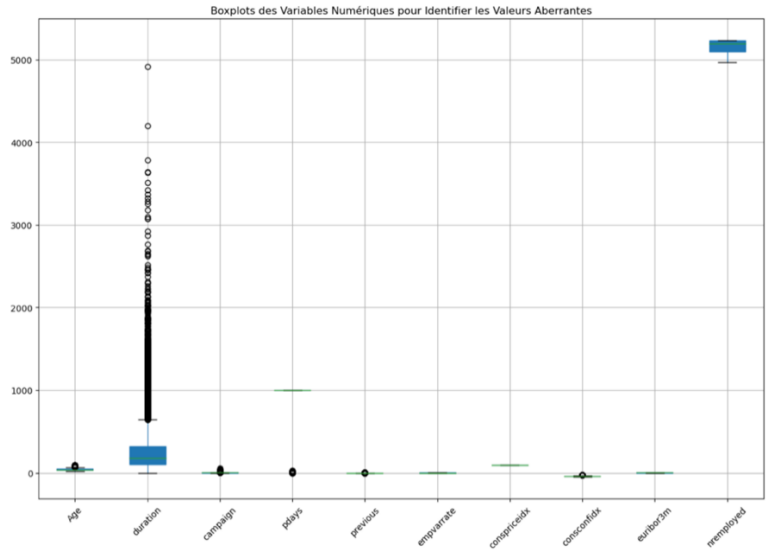
6. **Stratégie de Tarification** : Les résultats de cette analyse peuvent influencer la stratégie de tarification des produits de dépôt à terme. En ajustant les taux d'intérêt offerts sur les dépôts à terme en fonction des variations de l'Euribor, les banques peuvent maximiser l'attractivité de leurs offres et encourager plus de souscriptions.

Conclusion

En résumé, le graphique de la relation entre le taux Euribor à 3 mois et la souscription aux dépôts révèle que des taux Euribor plus bas sont associés à une plus grande probabilité de souscription à des dépôts à terme. Cette analyse offre des insights précieux pour les équipes marketing et de tarification, leur permettant de développer des stratégies adaptées aux conditions du marché pour optimiser les taux de souscription. En se basant sur ces données, les banques peuvent ajuster leurs offres pour répondre aux attentes des clients et améliorer leur compétitivité sur le marché financier.

3.2.2.8 Analyse des Valeurs aberrantes

Figure 19 : Boxplots des Variables Numériques pour Identifier les Valeurs aberrantes



Les principales observations du graphique sont les suivantes :

1. **Présence de Valeurs Extrêmes pour la Durée des Appels** : Le graphique boxplot montre que la variable "duration" (durée des appels) contient de nombreuses valeurs extrêmes, avec des durées allant jusqu'à plus de 5000 secondes. Cela suggère que certaines interactions téléphoniques sont exceptionnellement longues, ce qui pourrait nécessiter une enquête plus approfondie pour comprendre les raisons de ces longues durées.

2. **Variabilité des Variables Économiques** : La variable "nremployed" (nombre d'employés) montre également une variabilité notable, avec des valeurs s'étendant jusqu'à 5200. Cette variabilité peut refléter des différences significatives dans les contextes économiques au moment de la collecte des données, ce qui pourrait influencer les décisions de souscription.

3. **Variables avec Peu de Variabilité** : Les variables telles que "campaign" (nombre de contacts durant cette campagne), "pdays" (jours depuis le dernier contact de la campagne précédente), "previous" (nombre de contacts avant cette campagne), "empvarrate" (taux de variation de l'emploi), "conspriceidx" (indice des prix à la consommation) et "consconfidx" (indice de confiance des consommateurs) montrent peu de variabilité et peu de valeurs extrêmes. Cela suggère que ces variables sont relativement stables ou bien que les valeurs extrêmes ont été traitées efficacement.

4. **Identification des Outliers** : Les valeurs extrêmes (outliers) sont clairement identifiées pour chaque variable. Par exemple, la variable "duration" montre une large concentration de points au-delà de la plage normale, ce qui indique des durées d'appels typiquement longues. De même, "nremployed" présente quelques valeurs aberrantes qui peuvent être des cas extrêmes ou des erreurs de saisie.

5. **Implications pour le Nettoyage des Données** : La présence de valeurs extrêmes dans certaines variables, en particulier "duration", indique qu'un nettoyage supplémentaire des données peut être nécessaire. Par exemple, il peut être utile de traiter ou de supprimer ces valeurs extrêmes pour éviter qu'elles n'influencent négativement les résultats des analyses ou des modèles prédictifs.

6. **Stratégies d'Analyse** : Comprendre la distribution et la variabilité des variables permet de mieux préparer les données pour l'analyse. Les variables avec des valeurs extrêmes peuvent être transformées ou normalisées pour améliorer la performance des modèles prédictifs. De plus, les variables avec peu de variabilité peuvent être réévaluées pour déterminer leur pertinence dans l'analyse.

Conclusion

En résumé, les boxplots des variables numériques révèlent des informations cruciales sur la distribution des données et la présence de valeurs extrêmes. Cette analyse permet d'identifier les variables qui nécessitent un traitement particulier pour améliorer la qualité des données et la robustesse des analyses ultérieures. En traitant correctement les valeurs extrêmes et en normalisant les données, les banques peuvent améliorer la précision de leurs modèles prédictifs et optimiser leurs stratégies marketing.

3.2.2.9 Analyse des Valeurs Manquantes dans les Variables

Tableau 5 : Valeurs Manquantes dans les Variables

Age	0
Job	0
Marital	0
Education	0
Default	0
housing	0
Loan	0
Contact	0
Month	0
day_of_week	0
duration	0
campaign	0
pdays	0
previous	0
poutcome	0
empvarrate	0
conspriceidx	0
consconfidx	0
euribor3m	0
nremployed	0
y	0
dtype: int64	

Les principales observations du graphique sont les suivantes :

1. **Absence de Valeurs Manquantes** : Le tableau montre le nombre de valeurs manquantes pour chaque variable du dataset. Tous les champs indiquent une valeur de 0, ce qui signifie qu'il n'y a

pas de valeurs manquantes dans ce dataset. Chaque colonne dispose de toutes les données nécessaires pour chaque observation.

2. **Qualité des Données** : L'absence de valeurs manquantes indique une bonne qualité des données, ce qui simplifie le processus d'analyse et de modélisation. Les algorithmes de machine learning peuvent être appliqués directement sans nécessiter d'étapes supplémentaires de traitement des valeurs manquantes.

3. **Préparation des Données** : Avec des données complètes, la phase de préparation des données sera plus rapide et plus efficace. Les efforts peuvent être concentrés sur d'autres aspects, tels que la transformation des variables catégorielles en variables numériques, la normalisation des variables numériques, et le traitement des valeurs extrêmes.

4. **Fiabilité des Résultats** : Étant donné que le dataset est complet, les résultats des analyses et des modèles prédictifs seront plus fiables. Il n'y a pas de risque de biais introduit par l'imputation des valeurs manquantes ou par la suppression de lignes incomplètes.

5. **Facilité d'Exploration** : L'exploration des données et l'identification des tendances seront plus directes, car toutes les données nécessaires sont présentes. Cela permet une analyse exploratoire des données (EDA) plus exhaustive et plus précise.

6. **Optimisation des Modèles** : L'absence de valeurs manquantes permet d'utiliser pleinement les algorithmes de machine learning sans ajustements spécifiques pour gérer les données manquantes. Cela peut conduire à une meilleure performance des modèles prédictifs et à une optimisation plus efficace des hyperparamètres.

Conclusion

En résumé, l'absence de valeurs manquantes dans le dataset est un atout majeur pour l'analyse. Elle garantit une haute qualité des données et simplifie la préparation et la modélisation des données. Les résultats des analyses seront plus fiables et précis, permettant de tirer des insights plus pertinents pour les stratégies marketing de la banque. Les efforts peuvent être concentrés sur l'amélioration des modèles prédictifs et l'optimisation des campagnes marketing, en tirant parti des données complètes et de leurs intégrités.

3.3 Data Preparation

La préparation des données inclut plusieurs étapes : nettoyage des données, transformation des variables, traitement des valeurs manquantes, et préparation des données pour l'analyse ou la modélisation.

Voici les étapes que je vais suivre pour l'analyse :

- **Nettoyage des données** qui inclut la gestion des valeurs manquantes ainsi que la gestion des duplicatas

- **Transformation des variables** cette étape consiste à encoder des variables catégorielles ainsi que normaliser et standardiser des variables numériques si nécessaire.

- **Sélection des variables**: Sélectionner les variables les plus pertinentes pour le modèle.

3.3.1 Nettoyage des données :

Comme vu dans l'étape précédente, le dataset ne présente aucune valeur manquante, cependant il contient des duplicata. En effet, après exécution du code, 12 duplicatas ont été identifiés, ceux-ci ont été supprimés par la suite. Il est essentiel d'identifier les duplicatas c'est-à-dire des lignes identiques qui pourraient fausser les résultats de notre analyse. En identifiant et en supprimant les duplicatas, nous réduisons les risques de biais et d'erreurs dans les analyses ultérieures et les modèles prédictifs

3.3.2 Transformation des variables :

Dans cette section, je vais détailler les étapes de préparation des données. Je vais utiliser deux techniques principales : l'encodage one-hot pour les variables catégorielles et la standardisation pour les variables numériques.

3.3.3 Encodage One-Hot et Standardisation des Variables

Les variables catégorielles dans notre dataset contient des valeurs discrètes non ordinales, telles que des professions ou des niveaux d'éducation, qui ne peuvent pas être directement utilisées par les modèles de machine learning. C'est pourquoi afin de rendre ces informations compréhensibles directement par les modèles de machine learning je vais appliquer l'encodage one -hot. L'encodage one-hot crée une nouvelle colonne pour chaque catégorie unique de la variable catégorielle. Pour chaque observation, un '1' est placé dans la colonne correspondant à la catégorie présente et '0' dans toutes les autres colonnes. Par exemple, une variable "Job" avec les valeurs ["admin.", "blue-collar", "technician"] sera transformée en trois nouvelles colonnes : "Job_admin.", "Job_blue-collar", et "Job_technician".

Après l'application de l'encodage one-hot, les variables catégorielles sont transformées en plusieurs colonnes binaires, une pour chaque catégorie unique, ce qui permet aux algorithmes de machine learning de traiter ces données de manière efficace.

Une fois l'encodage one-hot terminé, il est important de standardiser les variables numériques, c'est-à-dire mettre à l'échelle numérique pour qu'elles aient une moyenne de zéro et un écart-type de 1. En faisant ceux-ci cela nous garantit que toutes les variables contribuent de manière égale à l'analyse.

Après avoir effectué l'encodage one-hot et la standardisation on a les variables continues : Âge, duration, campaign, pdays, previous, empvrarrate, conspriceidx, consconfidx, euribor3m, nremployed qui sont de type float64. Ce type de données indique que les variables sont des nombres flottants (ou décimaux) à 64 bits.

Pour les variables Catégorielles on a

- **Job_XXX** : Indique le type de job du client. Par exemple, Job_blue-collar, Job_entrepreneur, etc.
- **Marital_XXX** : Indique l'état civil du client. Par exemple, Marital_married, Marital_single, etc.
- **Education_XXX** : Niveau d'éducation du client. Par exemple, Education_basic.6y, Education_university.degree, etc.
- **Default_XXX** : Statut de défaut de paiement. Par exemple, Default_yes, Default_unknown.
- **housing_XXX** : Indique si le client a un prêt immobilier. Par exemple, housing_yes.
- **Loan_XXX** : Indique si le client a un prêt personnel. Par exemple, Loan_yes.
- **Contact_telephone** : Indique le moyen de contact, ici spécifiquement par téléphone.
- **Month_XXX** : Mois du dernier contact de la campagne. Par exemple, Month_aug, Month_dec, etc.
- **day_of_week_XXX** : Jour de la semaine du dernier contact. Par exemple, day_of_week_mon, day_of_week_thu, etc.
- **poutcome_XXX** : Résultat de la campagne de marketing précédente. Par exemple, poutcome_nonexistent, poutcome_success.

Le type de ces données est bool, ce type de données indique que les variables sont des booléens (True/False). Ces variables résultent de l'encodage one-hot qui a été effectué pour convertir les des variables catégorielles en variables binaires. Par exemple, pour la variable "Job", chaque type de job a été transformé en une colonne booléenne distincte indiquant la présence (True) ou l'absence (False) de ce type de job pour chaque observation.

3.3.4 Sélection des Variables

L'étape de sélection des caractéristiques est cruciale pour améliorer la performance et l'efficacité d'un modèle prédictif. Dans le cadre de cette analyse, j'ai utilisé un modèle de forêt aléatoire pour évaluer l'importance des différentes caractéristiques et sélectionner celles qui contribuent le plus à la prédiction de la souscription à un dépôt à terme.

La première étape consiste à identifier la colonne de notre ensemble de données qui contient l'information que nous voulons prédire. Cette colonne est appelée la "cible". Par exemple, si nous souhaitons prédire si un client va acheter un produit ou non, la colonne cible pourrait s'appeler "achat" et contenir des valeurs telles que "oui" ou "non". Définir clairement cette colonne est crucial, car elle sert de référence pour entraîner notre modèle prédictif. Dans notre cas cette colonne est appelée la colonne « y ». En effet c'est dans celle-ci que se trouve l'infirmation qui est de savoir si un client à souscrit ou non à un dépôt à terme.

Une fois cla colonne cible identifié, l'ensemble des données est séparé en deux parties distinctes : les caractéristiques et la cible. Les caractéristiques sont toutes les autres colonnes qui peuvent contenir des informations utilisées pour faire des prédictions. En effet, toutes les variables présentent dans le dataset ne sont pas utiles pour pouvoir répondre à la question qui est de savoir si un client va souscrire à un dépôt à terme ou non. C'est pourquoi j'utilise le modèle de forêt aléatoire qui va me permettre de déterminer quelles sont les caractéristiques à retenir pour le modèle prédictif.

Afin d'évaluer qu'elles sont les caractéristiques les plus pertinentes pour notre analyse, j'utilise le modèle de forêt aléatoire. Ainsi, chaque arbre de décision tente de prédire la cible en se basant sur différentes combinaisons de caractéristiques. En combinant les résultats de nombreux arbres, le modèle de forêt aléatoire peut fournir une estimation robuste de l'importance de chaque caractéristique. Nous entraînons

ce modèle en lui fournissant les caractéristiques (X) et la cible (y). Dans notre cas, la cible est une colonne nommée **y_yes**, qui indique si un client a souscrit à un dépôt à terme. Je prends cette colonne et la mets de côté, tandis que toutes les autres colonnes restent dans la partie caractéristique.

Une fois le modèle entraîné, j'utilise la technique du SelectFromModel qui permet de sélectionner les caractéristiques les plus importantes. En effet, elle analyse l'importance attribuée à chaque caractéristique par le modèle de forêt aléatoire et sélectionne celles qui dépassent une certaine moyenne. En d'autres termes, nous conservons uniquement les caractéristiques qui apportent une contribution significative à la prédiction de notre cible. Les caractéristiques sélectionnées après avoir appliqué le modèle de forêt aléaoitoire sont les suivantes : ('Age', 'duration', 'campaign', 'pdays', 'empvarrate', 'conspriceidx', 'consconfidx', 'euribor3m', 'nremployed', 'housing_yes', 'poutcome_success')

Enfin, un nouveau jeu de données est créé en utilisant uniquement les caractéristiques sélectionnées. Cela permet de réduire la complexité de notre modèle et de nous concentrer sur les informations les plus pertinentes. Ainsi cela nous permet de travailler avec un ensemble de données plus petit, mais de meilleure qualité qui améliore les performances de notre modèle prédictif, en rendant les prédictions plus précises et plus fiables.

3.4 Modeling

La phase de modélisation est cruciale pour développer des modèles prédictifs efficaces. Cette phase commence par le choix des techniques de modélisation appropriées. J'ai décidé d'appliquer plusieurs algorithmes dont : l'arbre de décision, le random forest, la régression et les réseaux de neurones.

Voici les étapes que je vais suivre pour cette étape :

- **Choix des techniques de modélisation**
- **Conception des tests**
- **Construction du modèle**
- **Évaluation du modèle** :

3.4.1 L'arbre de décision :

La première étape consiste à diviser les données en deux groupes : un groupe d'entraînement et un groupe de test. Le groupe d'entraînement est utilisé pour apprendre à l'algorithme à faire des prédictions, tandis que le groupe de test est utilisé pour vérifier si l'algorithme a bien appris. En général, nous utilisons 80% des données pour l'entraînement et 20% pour le test. Cela permet de vérifier que celui-ci peut faire des prédictions correctes sur des données qu'il n'a jamais vues auparavant.

Ensuite, j'ai utilisé un modèle d'arbre de décision. Le modèle regarde le groupe d'entraînement (les 80% de données) et apprend les relations entre les caractéristiques et la cible. En d'autres termes, le modèle apprend à reconnaître quels types de clients sont susceptibles de souscrire à un dépôt à terme.

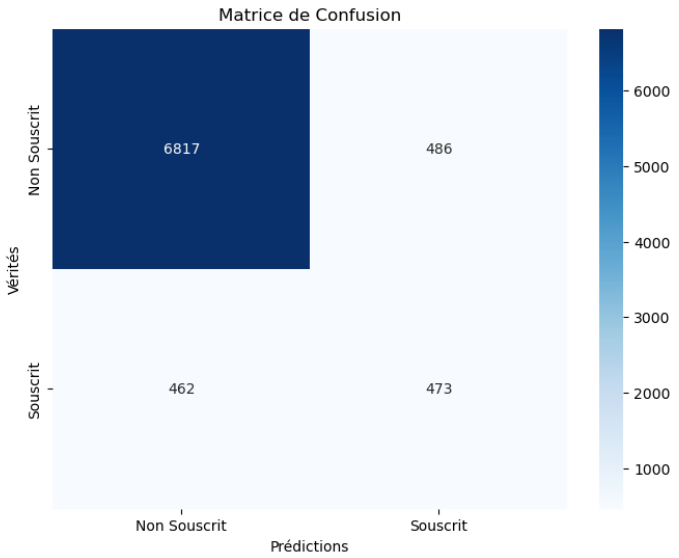
Enfin, après avoir appris à partir du groupe d'entraînement, j'ai demandé au modèle de faire des prédictions sur le groupe de test (les 20% de données restantes). C'est comme un examen pour vérifier si le modèle a bien appris. Les prédictions faites par le modèle sont ensuite comparées aux réponses réelles pour voir à quel point il a bien fonctionné.

Afin d'être sûr d'utiliser un modèle de qualité, j'ai évalué le modèle sur la base de plusieurs métriques, à savoir l'accuracy, la précision, le rappel et le F1-Score. Voici les résultats obtenus après l'évaluation du modèle. Ces métriques nous aident à comprendre comment le modèle performe.

Tableau 6 *:* Résumé de comparaison des métriques de performance : Arbre de décision

Métrique de performance	**Arbre de décision**
Accuracy	0,8849
Précision	0,4932
Rappel	0,5059
F1-Score	0,4995

Figure 20 : Matrice de Confusion Arbre de décision



- **Accuracy : 0,8849**

L'accuracy (cf. supra p.20), signifie que notre modèle a correctement prédit l'étiquette dans environ 88,49% des cas. En d'autres termes, sur 100 prédictions que fait notre modèle, environ 88 sont correctes.

C'est une mesure générale de la performance, mais elle peut parfois être trompeuse si nos données sont déséquilibrées.

- **Précision : 0,4932**

La précision (cf. supra p.20), mesure la qualité des prédictions positives de notre modèle. Ici, une précision de 49,32% signifie que lorsque notre modèle prédit qu'un client souscrira à un dépôt à terme, il a raison dans environ 49% des cas. C'est important pour comprendre la fiabilité des prédictions positives.

- **Rappel : 0,5059**

Le rappel (cf. supra p.20), mesure la capacité de notre modèle à identifier tous les clients qui souscriront effectivement à un dépôt à terme. Avec un rappel de 50,59%, cela signifie que notre modèle ne parvient à identifier qu'environ 50% des vrais cas de souscription. Un faible rappel indique que notre modèle manque une proportion significative de clients qui souscriront.

- **F1-Score : 0,4995**

Le F1-Score (cf. supra p.20), combine la précision et le rappel en une seule mesure. Avec un F1-Score de 0.4995, cela montre un équilibre modéré entre la précision et le rappel. C'est utile lorsque nous devons trouver un compromis entre les faux positifs et les faux négatifs.

- **Matrice de confusion**

La matrice de confusion (cf. supra p.23), donne une vue d'ensemble détaillée des performances de notre modèle en termes de vrais et faux positifs et négatifs :

- 6817 : Ce chiffre représente le nombre de cas où notre modèle a correctement prédit qu'un client ne souscrirait pas à un dépôt à terme (vrais négatifs).

- 486 : Ce chiffre représente le nombre de cas où notre modèle a incorrectement prédit qu'un client souscrirait à un dépôt à terme, mais ce n'était pas le cas (faux positifs).

- 462 : Ce chiffre représente le nombre de cas où notre modèle a incorrectement prédit qu'un client ne souscrirait pas à un dépôt à terme, alors qu'en réalité, il a souscrit (faux négatifs).

- 473 : Ce chiffre représente le nombre de cas où notre modèle a correctement prédit qu'un client souscrirait à un dépôt à terme (vrais positifs).

Conclusion

Les résultats montrent que notre modèle d'arbre de décision a une bonne précision globale (88,49%), ce qui signifie qu'il fait bien en général. Cependant, la précision et le rappel nous révèlent des aspects plus spécifiques : la précision de 49,32% est modérée, indiquant qu'environ la moitié des prédictions positives sont correctes. Le rappel de 50,59% montre que notre modèle identifie environ la moitié des clients qui souscriront effectivement à un dépôt à terme.

En conclusion, bien que le modèle d'arbre de décision montre une précision globale raisonnable, les métriques de précision et de rappel indiquent qu'il y a encore des améliorations à apporter, notamment en augmentant le rappel pour mieux identifier les souscriptions positives.

3.4.1.1 Amélioration du modèle (SMOTE)

L'une des principales difficultés dans l'apprentissage automatique survient lorsque les données sont déséquilibrées (cf. supra p.25), c'est-à-dire que les classes cibles ne sont pas représentées de manière égale. Par exemple, dans mon cas, le nombre de clients qui souscrivent à un dépôt à terme peut être beaucoup plus faible que ceux qui ne souscrivent pas. Ainsi un modèle entraîné sur des données déséquilibrées peut être biaisé en faveur de la classe majoritaire, réduisant ainsi la capacité du modèle à identifier correctement les éléments de la classe minoritaire. Pour remédier à cela, j'ai utilisé SMOTE (Synthetic Minority Over-sampling Technique), une méthode populaire pour générer des exemples synthétiques de la classe minoritaire afin d'équilibrer le dataset. Comme vu dans le chapitre 1, SMOTE est une méthode qui permet de créer des exemples synthétiques pour la classe minoritaire. SMOTE génère des exemples synthétiques pour la classe minoritaire en effectuant une interpolation entre les exemples existants de cette classe. Cela augmente le nombre d'exemples de la classe minoritaire, rendant le dataset plus équilibré. En faisant cela, SMOTE permet d'aider le modèle à mieux apprendre les caractéristiques de la classe minoritaire.

Pour commencer, j'ai de nouveau divisé les données en deux groupes : un groupe d'entraînement et un groupe de test. Comme dans l'étape précédente j'utilise 80% des données pour l'entraînement et 20% pour le test.

Après avoir appliqué SMOTE, j'ai entraîné un modèle d'arbre de décision sur les données équilibrées. Cela permet au modèle de mieux comprendre et prédire les exemples de la classe minoritaire.

Enfin, j'ai évalué le modèle en utilisant le groupe de test pour voir comment il performe sur des données qu'il n'a jamais vues auparavant. Voici les résultats obtenus après l'évaluation du modèle.

Tableau 7 : Résumé de comparaison des métriques de performance : Arbre de décision optimisé (SMOTE)

Métrique de performance	Arbre de décision	Arbre de décision optimisé (SMOTE)
Acurracy	0,8849	0,8824
Précison	0,4932	0,4852
Rappel	0,5059	0,5979
F1-Score	0,4995	0,5357

Avant l'application de SMOTE

Avant d'appliquer SMOTE, le modèle rencontrait des difficultés à prédire correctement les clients qui souscriraient à un dépôt à terme en raison du déséquilibre des classes. Cela se traduisait par un faible rappel et une précision modérée.

Après l'application de SMOTE

Après avoir appliqué SMOTE, le modèle a montré une meilleure capacité à prédire les souscriptions, comme le montre l'amélioration du rappel dans les résultats. Le F1-Score, qui combine la précision et le rappel, a également montré une amélioration, ce qui indique un meilleur équilibre global des performances du modèle.

Conclusion

L'application de SMOTE a permis d'améliorer significativement les performances du modèle en équilibrant les classes du dataset. Cela a conduit à une meilleure détection des souscriptions à un dépôt à terme, ce qui est crucial pour optimiser les stratégies marketing de la banque. En utilisant des données équilibrées, j'ai pu développer un modèle plus robuste et plus fiable pour prédire le comportement des clients.

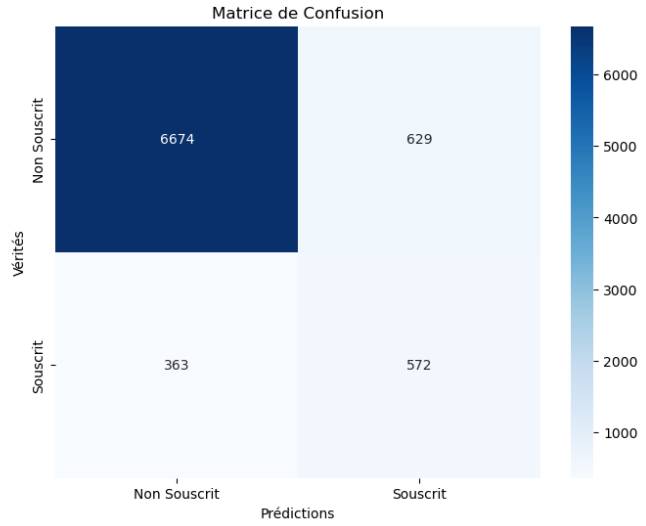
3.4.1.2 Amélioration du modèle (Optimisation des hyperparamètres)

En voulant continuer l'amélioration du modèle, je vais rajouter à la méthode SMOTE une autre technique qui est celle de l'optimisation des hyperparamètres (cf. supra p.26), avec GridSearchCV. Cette méthode peut améliorer encore plus les performances du modèle en trouvant la meilleure configuration possible. Voici les résultats obtenus après l'évaluation du modèle.

Tableau 8 : Résumé de comparaison des métriques de performance : Arbre de décision optimisé (SMOTE+GridSearchCV)

Métrique de performance	**Arbre de décision**	**Arbre de décision optimisé (SMOTE)**	**Arbre de décision optimisé (SMOTE+GridSearchCV)**
Acurracy	0,8849	0,8824	0,8796
Précison	0,4932	0,4852	0,4763
Rappel	0,5059	0,5979	0,6118
F1-Score	0,4995	0,5357	0,5356

Figure 21 : Matrice de Confusion : Arbre de décision optimisé (SMOTE+GridSearchCV)



Avec l'ajout de l'optimisation des hyperparamètres via GridSearchCV, le rappel a encore augmenté, indiquant une meilleure détection des clients susceptibles de souscrire à un dépôt à terme. Bien que la précision ait légèrement diminué, le compromis est acceptable, car le modèle est désormais plus robuste pour détecter les souscriptions. Le F1-Score reste similaire, montrant un bon équilibre entre la précision et le rappel.

Conclusion

Sans amélioration : Le modèle initial avait une bonne accuracy, mais la précision et le rappel étaient faibles, indiquant une mauvaise capacité à détecter les souscriptions.

Avec SMOTE : L'application de SMOTE a amélioré le rappel et le F1-Score, montrant une meilleure détection des classes minoritaires (souscriptions), malgré une légère diminution de la précision.

Avec SMOTE et optimisation des hyperparamètres : L'ajout de l'optimisation des hyperparamètres a encore amélioré le rappel tout en maintenant un bon équilibre général entre la précision et le rappel, rendant le modèle plus performant pour la détection des souscriptions.

En conclusion, l'optimisation des hyperparamètres combinée à SMOTE a permis de développer un modèle plus fiable et efficace pour prédire les comportements des clients, ce qui est essentiel pour optimiser les stratégies marketing de la banque.

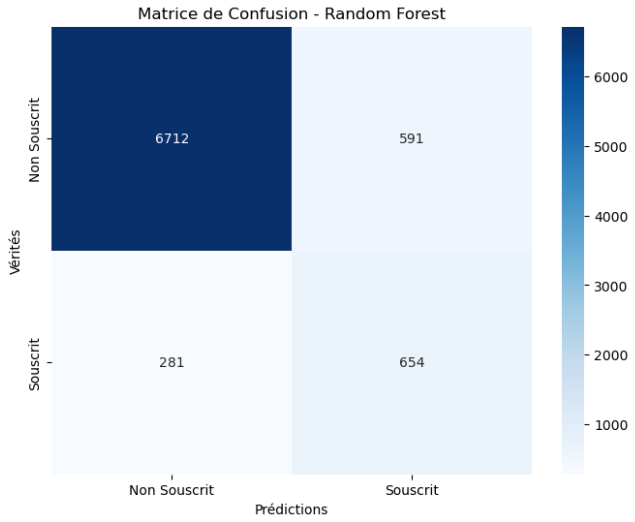
3.4.2 Random forest

Après avoir expérimenté avec le modèle de l'arbre de décision, notamment un arbre de décision avec SMOTE et une optimisation des hyperparamètres, j'ai voulu utiliser la méthode du Random Forest Classifier afin de pouvoir voir si les performances du modèle pouvaient être améliorées par rapport aux résultats obtenus avec le modèle d'arbre de décision le plus optimisé. Voici les résultats obtenus après l'évaluation du modèle

Tableau 9 : Résumé de comparaison des métriques de performance : Random Forest

Métrique de performance	Arbre de décision optimisé (SMOTE+GridSearchCV)	Random Forest
Acurracy	0,8796	0,8941
Précison	0,4763	0,5253
Rappel	0,6118	0,6995
F1-Score	0,5356	0,6000

Figure 22 : Matrice de Confusion : Random Forest



- **Accuracy : 0,8941**

Cela signifie que le modèle a correctement prédit l'étiquette dans environ 89,41% des cas. En d'autres termes, sur 100 prédictions faites par le modèle, environ 89 sont correctes. L'accuracy est une mesure générale de la performance, mais elle peut parfois être trompeuse si les données sont déséquilibrées.

- **Précision : 0,5253**

La précision mesure la qualité des prédictions positives de notre modèle. Ici, une précision de 52,53% signifie que lorsque le modèle prédit qu'un client souscrira à un dépôt à terme, il a raison dans environ 53% des cas. C'est important pour comprendre la fiabilité des prédictions positives.

- **Rappel : 0,6995**

Le rappel mesure la capacité du modèle à identifier tous les clients qui souscriront effectivement à un dépôt à terme. Avec un rappel de 69,95%, cela signifie que le modèle parvient à identifier environ 70% des vrais cas de souscription. Un rappel élevé indique que le modèle manque moins de clients susceptibles de souscrire.

- **F1-Score : 0,6000**

Le F1-Score combine la précision et le rappel en une seule mesure. Avec un F1-Score de 0,6000, cela montre un équilibre relativement bon entre la précision et le rappel. Cette mesure est utile lorsque nous devons trouver un compromis entre les faux positifs et les faux négatifs.

- **Matrice de confusion**

- 6712 : Ce chiffre représente le nombre de cas où le modèle a correctement prédit qu'un client ne souscrirait pas à un dépôt à terme (vrais négatifs).
- 591 : Ce chiffre représente le nombre de cas où le modèle a incorrectement prédit qu'un client souscrirait à un dépôt à terme, mais ce n'était pas le cas (faux positifs).

- 281 : Ce chiffre représente le nombre de cas où le modèle a incorrectement prédit qu'un client ne souscrirait pas à un dépôt à terme, alors qu'en réalité, il a souscrit (faux négatifs).

- 654 : Ce chiffre représente le nombre de cas où le modèle a correctement prédit qu'un client souscrirait à un dépôt à terme (vrais positifs).

Conclusion

Le Random Forest réduit le nombre de faux négatifs de 363 à 281, ce qui signifie qu'il identifie correctement 82 clients supplémentaires qui souscriront. De plus, il y a une légère diminution des faux positifs de 629 à 591.

Le modèle Random Forest surpasse l'arbre de décision optimisé dans presque tous les aspects critiques de la performance prédictive. Grâce à un rappel plus élevé, le Random Forest identifie plus efficacement les clients susceptibles de souscrire, ce qui est essentiel pour minimiser les pertes de revenus potentielles.

De plus, avec une précision plus élevée, les prédictions positives du Random Forest sont plus fiables, réduisant ainsi les efforts marketing inutiles dirigés vers les clients qui ne souscriront pas.

Ensuite, le F1-Score supérieur indique que le Random Forest maintient un bon équilibre entre la réduction des faux positifs et l'amélioration de la détection des vrais positifs.

En conclusion, l'adoption du Random Forest pour la prédiction des souscriptions à un dépôt à terme est fortement recommandée. Ce modèle offre une meilleure performance globale, permettant à la banque d'optimiser ses stratégies marketing et de cibler plus efficacement les clients potentiels.

3.4.3 La Régression

Je vais maintenant passer à l'application du modèle de régression. Comme pour les modèles précédents, je divise les données en deux groupes (80 % pour l'entraînement et 20 % pour le test). Cela me permet d'apprendre au modèle à partir des données d'entraînement et de vérifier sa performance sur des données qu'il n'a jamais vues auparavant. Cette méthode garantit une comparaison cohérente avec les autres modèles utilisés précédemment.

Ensuite, j'utilise un modèle de régression logistique. Je dis au modèle de regarder le groupe d'entraînement (les 80% de données) et d'apprendre les relations entre les caractéristiques et la cible. En d'autres termes, le modèle apprend à reconnaître quels types de clients sont susceptibles de souscrire à un dépôt à terme.

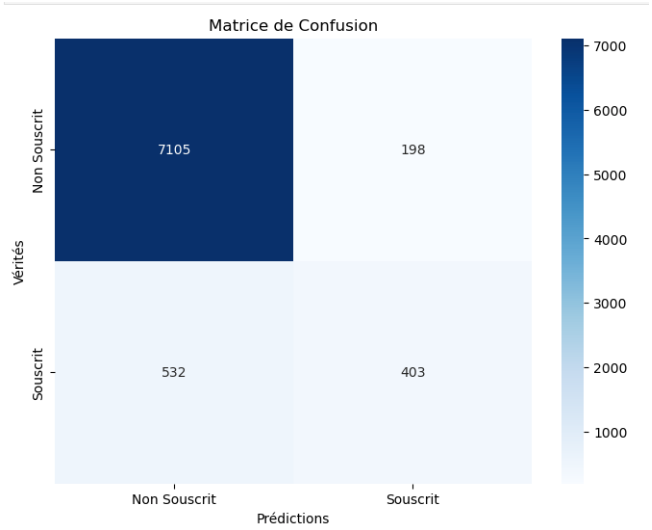
En outre, après avoir appris à partir du groupe d'entraînement, je demande au modèle de faire des prédictions sur le groupe de test (les 20% de données restantes). Les prédictions faites par le modèle sont ensuite comparées aux réponses réelles pour voir à quel point il a bien fonctionné.

Enfin, afin d'être sûr d'utiliser un modèle de qualité, je vais évaluer le modèle sur base des mêmes métriques utilisé pour évaluer les modèles précédents à savoir l'accuracy, la précision, le rappel et la F1-Score. Voici les résultats obtenus après l'évaluation du modèle. Ces métriques vont nous aider à comprendre comment le modèle performe.

Tableau 10 : Résumé des métriques de performance : Régression

Métrique de performance	**Régression**
Accuracy	0,9114
Précision	0,6705
Rappel	0,4310
F1-Score	0,5247

Figure 23 : Matrice de Confusion : Régression



- **Accuracy: 0,9114**

Cela signifie que notre modèle a correctement prédit l'étiquette dans environ 91,14% des cas. En d'autres termes, sur 100 prédictions que fait notre modèle, environ 91 sont correctes.

- **Précision : 0,6705**

La précision mesure la qualité des prédictions positives de notre modèle. Ici, une précision de 67,05% signifie que lorsque notre modèle prédit qu'un client souscrira à un dépôt à terme, il a raison dans environ 67% des cas. C'est important pour comprendre la fiabilité des prédictions positives.

- **Rappel : 0,4310**

Avec un rappel de 43,10%, cela signifie que notre modèle ne parvient à identifier qu'environ 43% des vrais cas de souscription. Un faible rappel indique que notre modèle manque une proportion significative de clients qui souscriront.

- **F1-Score : 0,5247**

Avec un F1-Score de 0,5247, cela montre un équilibre modéré entre la précision et le rappel. C'est utile lorsque nous devons trouver un compromis entre les faux positifs et les faux négatifs.

- **La matrice de confusion**

- 7105 : Ce chiffre représente le nombre de cas où notre modèle a correctement prédit qu'un client ne souscrirait pas à un dépôt à terme (vrais négatifs).

- 198 : Ce chiffre représente le nombre de cas où notre modèle a incorrectement prédit qu'un client souscrirait à un dépôt à terme, mais ce n'était pas le cas (faux positifs).

- 532 : Ce chiffre représente le nombre de cas où notre modèle a incorrectement prédit qu'un client ne souscrirait pas à un dépôt à terme, alors qu'en réalité, il a souscrit (faux négatifs).

- 403 : Ce chiffre représente le nombre de cas où notre modèle a correctement prédit qu'un client souscrirait à un dépôt à terme (vrais positifs).

Conclusion

Les résultats montrent que notre modèle de régression logistique a une bonne précision globale (91,14%). Cependant, la précision et le rappel nous révèlent des aspects plus spécifiques :

La précision de 67,05% est relativement bonne, indiquant que la majorité des prédictions positives sont correctes.

Le rappel de 43,10% est plus préoccupant, car il montre que notre modèle ne parvient pas à identifier une grande partie des clients qui souscriront effectivement à un dépôt à terme.

En conclusion, bien que le modèle de régression logistique montre une précision globale élevée, les métriques de précision et de rappel indiquent qu'il y a encore des améliorations à apporter, notamment en augmentant le rappel pour mieux identifier les souscriptions positives. Cela pourrait être amélioré par des techniques d'optimisation de modèle, telles que l'équilibrage des classes ou l'ajout de nouvelles fonctionnalités.

3.4.4 Customer2Vec

Après avoir expérimenté avec divers modèles, notamment un arbre de décision avec SMOTE et une optimisation des hyperparamètres, ainsi qu'un modèle Random Forest et la Régression j'ai constaté que l'utilisation d'un modèle Customer2Vec pourrait potentiellement améliorer encore plus les performances prédictives.

Comme vu (cf. supra p.13), Customer2Vec est une technique qui transforme les comportements des clients en vecteurs denses, similaires aux techniques utilisées dans les modèles Word2Vec pour le traitement du langage naturel. Cette approche permet de capturer les complexités des comportements des clients et de représenter chaque client comme un vecteur dense dans un espace de caractéristiques de dimension réduite.

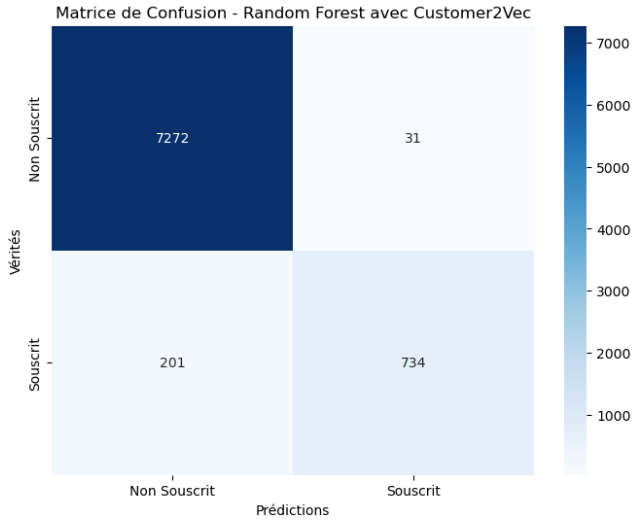
Un vecteur d'embedding a été créé pour chaque client. Ces vecteurs ont été utilisés comme caractéristiques pour entraîner un modèle de Random Forest. En capturant les relations complexes et en créant des représentations denses des clients. On fournit des informations précieuses qui peuvent améliorer la performance des modèles prédictifs.

Voici les résultats obtenus après la combinaison de Customer2Vec et Random Forest :

Tableau 11 : Résumé des métriques de performance : Random Forest avec Customer2Vec

Métrique de performance	Arbre de décision	Arbre de décision optimisé (SMOTE)	Arbre de décision optimisé (SMOTE+GridSearchCV)	Random Forest	Radam Forest avec Customer2Vec
Acurracy	0,8849	0,8824	0,8796	0,8941	0,9718
Précison	0,4932	0,4852	0,4763	0,5253	0,9595
Rappel	0,5059	0,5979	0,6118	0,6995	0,7850
F1-Score	0,4995	0,5357	0,5356	0,6000	0,8635

Figure 24 : Matrice de confusion : Random Forest avec Customer2Vec



- **Accuracy : 0,9718**

Cela signifie que mon modèle a correctement prédit l'étiquette dans environ 97,18% des cas. En d'autres termes, sur 100 prédictions faites par mon modèle, environ 97 sont correctes.

- **Précision : 0,9595**

Ici on a une précision de 95,95% signifie que lorsque mon modèle prédit qu'un client souscrira à un dépôt à terme, il a raison dans environ 96% des cas.

- **Rappel : 0,7850**

Le rappel mesure la capacité de mon modèle à identifier tous les clients qui souscriront effectivement à un dépôt à terme. Avec un rappel de 78,50%, cela signifie que mon modèle parvient à identifier environ 79% des vrais cas de souscription.

- **F1-Score : 0,8635**

Le F1-Score combine la précision et le rappel en une seule mesure. Avec un F1-Score de 0.8635, cela montre un excellent équilibre entre la précision et le rappel.

- **Matrice de confusion**

- 7272 : Ce chiffre représente le nombre de cas où mon modèle a correctement prédit qu'un client ne souscrirait pas à un dépôt à terme (vrais négatifs).

- 31 : Ce chiffre représente le nombre de cas où mon modèle a incorrectement prédit qu'un client souscrirait à un dépôt à terme, mais ce n'était pas le cas (faux positifs).

- 201 : Ce chiffre représente le nombre de cas où mon modèle a incorrectement prédit qu'un client ne souscrirait pas à un dépôt à terme, alors qu'en réalité, il a souscrit (faux négatifs).

- 734 : Ce chiffre représente le nombre de cas où mon modèle a correctement prédit qu'un client souscrirait à un dépôt à terme (vrais positifs).

Conclusion

Le modèle Random Forest avec Customer2Vec affiche une précision de 95,95%, montrant une capacité exceptionnelle à éviter les faux positifs. Le rappel de 78,50% indique que le modèle identifie efficacement la majorité des clients susceptibles de souscrire, et le F1-Score de 86,35% montre un excellent équilibre entre précision et rappel.

L'approche Customer2Vec combinée avec un modèle Random Forest s'est révélée être une méthode extrêmement efficace pour prédire les souscriptions à un dépôt à terme. Cette approche permet non seulement de capturer les complexités des comportements des clients grâce aux embeddings denses, mais aussi d'exploiter la robustesse et la performance des modèles d'ensemble comme Random Forest.

3.5 Sélection des modèles

Après avoir évalué divers modèles prédictifs, notamment des arbres de décision optimisés avec SMOTE et GridSearchCV, ainsi que des modèles de Random Forest avec et sans Customer2Vec, j’ai décidé

de garder deux modèles pour le déploiement final. Le premier modèle sélectionné est le **Random Forest avec Customer2Vec**, qui a démontré des performances exceptionnelles avec une précision de 95,95%, un rappel de 78,50%, et un F1-Score de 86,35%. Ce modèle excelle dans la capture des complexités des comportements clients grâce à l'utilisation de vecteurs denses, ce qui permet un ciblage précis et une personnalisation des offres.

Le deuxième modèle retenu est le **Random Forest**, qui a montré une robustesse et un équilibre notable avec une précision de 52,53%, un rappel de 69,95%, et un F1-Score de 60,00%. Ce modèle peut servir de modèle de secours ou de comparaison pour valider les performances du modèle principal. En intégrant ces deux modèles, nous visons à optimiser les campagnes marketing en ciblant plus efficacement les clients potentiels et en personnalisant les offres pour améliorer la satisfaction des clients.

3.6 Déploiement

Le modèle DELTA Plus, décrit précédemment (cf. supra p. 35), fournit un cadre pour évaluer la maturité analytique d'une organisation. Cette maturité est cruciale pour réussir le déploiement de modèles prédictifs dans des applications commerciales concrètes. Par exemple, le déploiement d'un modèle prédictif tel que celui développé ici consiste à l'utiliser de manière systématique dans des campagnes marketing pour un produit financier, comme un prêt bancaire. Le ciblage résultant du modèle prédictif permet de segmenter les clients de manière fine et de personnaliser les offres, augmentant ainsi les chances de succès des campagnes.

La capacité d'une organisation à développer et à déployer efficacement de tels modèles prédictifs est un indicateur clé de sa maturité analytique. Le modèle DELTA+ fournit un cadre de réflexion pour évaluer cette maturité, en particulier à travers des composantes telles que les Données (Data), les Techniques Analytiques Avancées (Advanced Analytics) et les Cibles (Targets). Ces éléments sont essentiels pour comprendre comment une organisation peut tirer parti de l'analytique dans le cadre de campagnes marketing ciblées et comment cette approche peut être accueillie par l'organisation dans son ensemble.

Data : L'utilisation de données de haute qualité est indispensable pour le bon fonctionnement des modèles prédictifs. Les données comportementales et démographiques que j'ai utilisées pour cibler les clients lors des campagnes marketing illustrent cette exigence. Ces données permettent une personnalisation efficace des offres et alignent le processus de modélisation avec les principes du modèle DELTA Plus.

Enterprise : Le déploiement systématique de modèles prédictifs dans l'organisation reflète une orientation analytique centralisée. Cela montre que l'entreprise est prête à progresser vers des niveaux de maturité supérieurs, tels que "Analytical Aspirations" et "Analytical Companies", où les analyses sont utilisées de manière coordonnée à travers l'organisation.

Leadership : Le rôle du leadership est crucial pour soutenir l'adoption des modèles prédictifs dans une organisation. L'engagement des dirigeants à promouvoir l'usage des technologies avancées et des compétences analytiques est un facteur déterminant pour la réussite du déploiement. Si le leadership soutient pleinement ces initiatives, l'organisation pourra tirer un avantage compétitif grâce à l'utilisation systématique de l'analytique.

Targets : Le déploiement d'un modèle prédictif pour cibler des segments spécifiques de clients, comme dans le cadre des campagnes marketing personnalisées, s'aligne avec l'un des objectifs stratégiques du modèle DELTA Plus. En appliquant ces analyses pour personnaliser les offres, l'organisation atteint des cibles stratégiques et exploite pleinement ses ressources analytiques pour maximiser l'efficacité des campagnes.

Analysts et Advanced Analytics : Le développement des compétences analytiques, notamment avec des outils comme Random Forests et Customer2Vec, est un indicateur de progression vers une plus grande maturité analytique. Ces techniques permettent une analyse sophistiquée des comportements clients et contribuent à l'amélioration des modèles prédictifs, rendant l'organisation plus apte à utiliser ces informations de manière stratégique.

En conclusion, la structure et les concepts que j'ai développés dans cette phase exploratoire de mon mémoire reflètent une adoption progressive des éléments du DELTA Plus Model. Cela permet à l'organisation de progresser vers des niveaux plus avancés de maturité analytique, garantissant ainsi une utilisation plus efficace et compétitive des données.

3.6.1 Expérimentation Commerciale : Impact des Campagnes Marketing Personnalisées sur la Souscription à des Prêts Bancaires

Dans le cadre de mon mémoire, j'explore comment l'analyse des données peut être utilisée pour cibler efficacement les clients et personnaliser les offres dans les campagnes marketing des banques. L'objectif principal est d'identifier les segments de clients les plus susceptibles de souscrire à des produits financiers, comme les prêts, suite à des campagnes marketing ciblées. Si cette expérimentation devait être menée en conditions réelles, elle me permettrait d'évaluer précisément l'impact de la personnalisation sur les décisions des clients.

Hypothèse :

Je suppose que les clients exposés à une campagne marketing personnalisée pour des prêts seront plus enclins à souscrire à un prêt par rapport à ceux qui reçoivent une campagne marketing générique.

Plan d'Expérimentation :

Si je devais mettre en place cette expérimentation sur le terrain, voici comment je m'y prendrais.

Objectif de l'Expérience :

Déterminer l'efficacité des campagnes marketing personnalisées comparées aux campagnes génériques pour encourager la souscription à des prêts. Cela impliquerait de tester directement l'hypothèse en exposant différents groupes de clients à des campagnes distinctes (personnalisée et générique).

Conception de l'Expérience :

Groupe A (test) : Ce groupe recevrait une campagne marketing personnalisée, basée sur leurs données comportementales et démographiques. Par exemple, des offres de prêt avec des taux d'intérêt préférentiels ou des messages sur-mesure seraient envoyés en fonction de leur profil.

Groupe B (contrôle) : Ce groupe recevrait une campagne marketing générique, sans personnalisation particulière. Les offres seraient standardisées et identiques pour tous les clients du groupe.

Si l'expérimentation devait être réalisée, il serait crucial de définir précisément la méthode de randomisation des clients dans ces deux groupes pour garantir l'intégrité des résultats. Cela permettrait d'assurer que les différences observées dans les taux de souscription sont bien liées à la personnalisation des campagnes.

Variables :

Les variables à analyser dans cette expérimentation seraient :

- **Variable indépendante :** Le type de campagne marketing (personnalisée vs. générique), qui déterminerait l'impact des offres spécifiques sur le comportement des clients.

- **Variable dépendante :** Le taux de souscription à un prêt, mesuré en fonction des interactions des clients avec les campagnes.

Collecte de Données :

Si je devais collecter les données, voici comment cela se passerait : je mesurerais le taux de souscription à un prêt pour chaque groupe, en suivant également les interactions des clients avec la campagne (ouverture des emails, clics, réponses aux offres, etc.). Cela permettrait de quantifier précisément l'impact de la personnalisation sur le comportement des clients.

Pourquoi cette approche fonctionnerait en pratique ?

L'expérimentation que je décris ici suit les principes rigoureux d'une méthodologie scientifique, tels que décrits par Thomke et Manzi (2014), dans "The Discipline of Business Experimentation". Si cette expérimentation devait être mise en œuvre, elle respecterait les mêmes standards :

- Hypothèse claire et testable : L'hypothèse testée ici est simple et mesurable, ce qui facilite l'analyse des résultats.

- Groupes test et contrôle définis : La création de deux groupes, test et contrôle, permettrait de comparer les résultats de manière significative.

- Collecte de données fiable : En recueillant les données d'interaction des clients avec les campagnes (taux de clics, souscriptions, etc.), je m'assurerais que l'expérimentation est bien fondée sur des mesures précises et tangibles.

En suivant cette méthodologie, je pourrais valider ou invalider l'hypothèse selon laquelle les campagnes marketing personnalisées sont plus efficaces que les campagnes génériques en termes de taux de souscription.

3.6.2 Segmentation des clients et déploiement des campagnes marketing

Dans le cadre de mon expérimentation commerciale visant à optimiser les stratégies de vente et les offres financières, je réalise une segmentation des clients en utilisant l'algorithme de K-Means Clustering. Cette segmentation me permet d'identifier des groupes de clients partageant des caractéristiques similaires, facilitant ainsi la mise en place de campagnes marketing plus ciblées et pertinentes.

1. Méthodologie de segmentation

Je mène l'analyse sur un jeu de données comprenant des variables sociodémographiques (comme l'âge, l'emploi, le statut marital) ainsi que des variables économiques (par exemple le taux Euribor, les emprunts). Grâce à cette segmentation, j'identifie trois clusters distincts, chacun ayant ses propres traits caractéristiques et des comportements différents face aux offres de prêts proposées.

- **Cluster 0 (Violet)** : Ce groupe se compose principalement de clients âgés de 30 à 60 ans, avec des taux d'Euribor relativement bas, généralement autour de 1 à 2 %. Les membres de ce groupe ont tendance à ne pas souscrire massivement aux offres de prêt, bien que leurs revenus semblent leur permettre de prendre des décisions financières stables.

- **Cluster 1 (Jaune)** : Ce segment regroupe des clients principalement âgés de 30 à 50 ans, ayant des taux Euribor élevés (entre 4 et 5 %). Ces clients se montrent plus enclins à souscrire aux offres bancaires proposées, ce qui est cohérent avec leur profil : des revenus probablement plus élevés et une maturité financière plus grande. Il s'agit du groupe présentant la plus forte probabilité de conversion.

- **Cluster 2 (Vert)** : Ce groupe est composé majoritairement de jeunes clients (moins de 40 ans), ayant des taux Euribor relativement bas, souvent en dessous de 1 %. Ce segment montre un faible engagement envers les offres proposées, ce qui peut s'expliquer par des priorités financières différentes ou une aversion aux prêts à long terme.

2. Analyse des taux de conversion par segment

Mon objectif principal avec cette segmentation est d'identifier les segments les plus réactifs aux campagnes marketing, afin d'optimiser les ressources allouées aux différentes actions. Pour ce faire, je calcule le taux de conversion, c'est-à-dire la proportion de clients ayant souscrit à une offre, pour chaque cluster.

Les résultats obtenus sont les suivants :

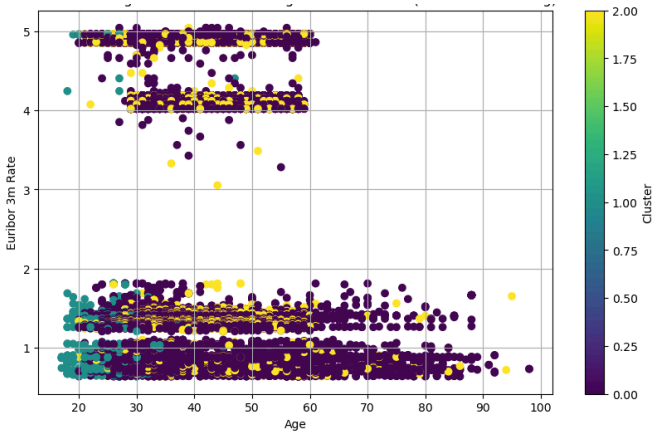
- **Cluster 0 (Violet)** : Taux de conversion de 12,48 %
- **Cluster 1 (Jaune)** : Taux de conversion de 31,43 % (le plus élevé)
- **Cluster 2 (Vert)** : Taux de conversion de 6,99 % (le plus faible)

Ces résultats montrent clairement que le Cluster 1 est le plus prometteur en termes de conversion, avec plus de 31 % des clients souscrivant à l'offre. Ce segment, qui regroupe des clients d'âge intermédiaire avec des taux Euribor élevés, est probablement composé de clients ayant une plus grande capacité financière ou un besoin accru de produits bancaires, justifiant ainsi une meilleure réponse aux offres. En revanche, le Cluster 2, bien qu'il soit majoritairement composé de jeunes clients, semble moins réceptif aux offres, probablement en raison de leur stade de vie ou de leurs besoins financiers immédiats.

3. Visualisation et interprétation des segments

Le graphique ci-dessous illustre la segmentation des clients selon l'âge et le taux Euribor à 3 mois, offrant une meilleure compréhension visuelle des différents groupes identifiés :

Figure 25 : Segmentation des clients selon l'âge et le taux Euribor à 3 mois



Ce graphique montre que le Cluster 1 (Jaune) se concentre autour de tranches d'âge intermédiaires (30-50 ans) avec des taux Euribor élevés, ce qui peut expliquer leur plus grande propension à souscrire aux offres bancaires. À l'inverse, le Cluster 2 (Vert) regroupe principalement des clients plus jeunes, qui semblent être moins enclins à souscrire.

4. Implications pour les campagnes marketing

Les résultats de cette analyse fournissent des indications claires sur les stratégies à adopter pour chaque segment.

- **Cluster 1 (Jaune)** : Ce segment, ayant un taux de conversion élevé de 31,43 %, représente une opportunité clé pour des campagnes marketing ciblées. Les campagnes peuvent être optimisées en se concentrant sur des offres de crédit à moyen ou long terme, adaptées aux besoins de cette

tranche d'âge, qui est probablement plus disposée à souscrire à des emprunts avec des taux d'intérêt plus élevés.

- **Cluster 0 (Violet) :** Bien que le taux de conversion de ce cluster soit inférieur à celui du Cluster 1, il reste intéressant avec un taux de 12,48 %. Des campagnes de fidélisation ou d'incitation peuvent être envisagées pour ce groupe, en leur proposant des taux plus compétitifs ou des produits complémentaires.

- **Cluster 2 (Vert) :** Ce groupe, ayant un taux de conversion de seulement 6,99 %, nécessite une approche différente. Étant principalement composé de jeunes clients, des offres alternatives ou des produits bancaires adaptés à un public plus jeune (par exemple, des prêts à court terme ou des solutions d'épargne) peuvent mieux répondre à leurs besoins.

5. Recommandations stratégiques basées sur la segmentation

Sur la base des analyses effectuées, voici les recommandations stratégiques que je propose :

Focaliser les ressources marketing sur le Cluster 1. Ce groupe présente la plus forte probabilité de conversion, avec des clients ayant une maturité financière et une plus grande capacité à souscrire aux offres. Il est donc essentiel d'adapter des offres spécifiques pour maximiser le taux de souscription.

Adapter les offres pour le Cluster 0. En proposant des produits plus personnalisés, tels que des réductions sur les taux d'intérêt ou des facilités de remboursement, il est possible d'améliorer le taux de conversion de ce groupe, qui reste relativement prometteur.

Réorienter la stratégie pour le Cluster 2. Ce groupe est actuellement moins réceptif aux offres actuelles. Je recommande de mener des études supplémentaires pour mieux comprendre les besoins spécifiques de ces clients et d'identifier des produits plus adaptés à leur profil (par exemple, des solutions d'épargne à court terme ou des prêts à taux réduit).

4 Prise de Recul

Dans le cadre de cette recherche, l'analyse des données s'est révélée être une méthode puissante pour cibler efficacement les clients et personnaliser les offres marketing. Toutefois, il est crucial de reconnaître que notre dataset présente certaines limitations qui peuvent influencer les résultats obtenus et les conclusions de l'étude.

Premièrement, la qualité des données joue un rôle fondamental dans l'analyse. Des données manquantes, incomplètes ou incorrectes peuvent biaiser les résultats. Dans notre dataset, certaines variables, comme Default, comportaient des valeurs « unknown » qui nécessitaient des étapes de nettoyage et d'imputation des données. Bien que des techniques de correction aient été appliquées, il est possible que des biais persistent. Deuxièmement, le dataset utilisé pourrait ne pas être entièrement représentatif de l'ensemble de la population de clients de la banque. Si l'échantillon est biaisé, par exemple s'il inclut principalement des clients d'une certaine région ou d'un certain groupe démographique, les résultats de l'analyse pourraient ne pas être applicable à toute la base de clients de la banque.

En outre, les données utilisées dans cette étude sont statiques et représentent une période spécifique, comme le mois de mai dans la variable « Month ». Les comportements des clients et les tendances du marché peuvent évoluer avec le temps, ce qui limite la validité des conclusions à long terme. Des analyses longitudinales avec des données mises à jour seraient nécessaires pour suivre les évolutions des comportements des clients.

Par ailleurs, le dataset se limite aux variables disponibles au moment de la collecte des données. Certaines informations pertinentes pour une segmentation ou une personnalisation plus fine des offres pourraient manquer. Par exemple, des données sur les interactions en temps réel avec les services numériques de la banque ou des informations qualitatives sur la satisfaction des clients pourraient enrichir l'analyse.

Enfin, les modèles prédictifs utilisés, tels que l'arbre de décision ou le modèle Customer2Vec, peuvent introduire des biais si les données d'entraînement sont biaisées. De plus, ces modèles peuvent être sensibles aux hyperparamètres choisis et aux techniques de prétraitement des données, ce qui peut affecter leurs performances et leur interprétabilité.

5 Recommandations

Pour atténuer ces limitations et renforcer la robustesse des futures recherches, plusieurs recommandations peuvent être envisagées.

Premièrement, il serait bénéfique de mettre en place des processus rigoureux de collecte et de validation des données pour minimiser les erreurs et les valeurs manquantes. L'intégration de sources de données multiples peut également enrichir la base de données.

Deuxièmement, il est crucial de s'assurer que l'échantillon de données est représentatif de l'ensemble de la population de clients.

Troisièmement, la collecte de données sur plusieurs périodes permettrait d'analyser les tendances et les évolutions des comportements des clients dans le temps.

Quatrièmement, il serait avantageux d'inclure des variables additionnelles pertinentes pour l'analyse, telles que des données sur les interactions en temps réel et des feedbacks qualitatifs des clients.

Enfin, il est important de continuer à tester et à évaluer les modèles prédictifs avec différentes techniques et paramètres pour améliorer leur précision et leur interprétabilité. Considérer l'utilisation de techniques d'apprentissage automatique avancées permettrait de mieux capturer les nuances des comportements des clients.

En reconnaissant et en adressant ces limitations, cette recherche vise à offrir une base pour des analyses futures et contribue à l'amélioration continue des stratégies de marketing bancaire basées sur l'analyse des données.

6 Conclusion

En conclusion, cette étude a mis en lumière l'importance cruciale de l'analyse des données dans le secteur bancaire, particulièrement en ce qui concerne le ciblage des clients et la personnalisation des offres marketing. À travers une analyse rigoureuse des comportements des clients, il a été possible de démontrer comment les banques peuvent tirer parti des données pour améliorer significativement l'efficacité de leurs campagnes marketing.

La revue de la littérature a permis de situer notre étude dans le contexte plus large des évolutions technologiques et des pratiques marketing dans le secteur bancaire. Elle a révélé que l'adoption de technologies numériques et l'exploitation de données volumineuses sont devenues des facteurs déterminants pour le succès des banques modernes. Les concepts de segmentation de la clientèle et d'offres personnalisées ont été explorés en profondeur, mettant en évidence l'importance de ces éléments pour répondre aux attentes des clients et améliorer leur satisfaction.

Ensuite, l'étude empirique que j'ai réalisée a démontré que l'utilisation de modèles prédictifs, tels que l'arbre de décision et Customer2Vec, permet de mieux comprendre les préférences et les besoins des clients, offrant ainsi une base solide pour développer des offres personnalisées. Cependant, il est essentiel de reconnaître et d'aborder les limitations inhérentes au dataset utilisé. La qualité des données, la représentativité de l'échantillon, la temporalité des données et les variables disponibles sont autant de facteurs qui peuvent influencer les résultats et leur interprétation.

Pour renforcer la robustesse des analyses futures, plusieurs recommandations ont été proposées. Il est crucial de mettre en place des processus rigoureux de collecte et de validation des données, d'assurer un échantillonnage représentatif de la population de clients, de collecter des données longitudinales, d'enrichir le dataset avec des variables additionnelles pertinentes et de continuer à tester et évaluer les modèles prédictifs avec différentes techniques et paramètres.

En somme, cette recherche contribue à une meilleure compréhension de l'impact de l'analyse des données sur les stratégies marketing des banques. En intégrant les recommandations proposées, les banques peuvent améliorer continuellement leurs approches et développer des campagnes marketing plus ciblées et efficaces, répondant ainsi aux attentes de leurs clients dans un environnement de plus en plus concurrentiel et digitalisé.

La prise de conscience des limites du dataset et la mise en œuvre des améliorations suggérées permettront non seulement de renforcer la validité des analyses, mais aussi de maximiser l'impact des stratégies marketing sur la satisfaction et la fidélisation des clients. Ainsi, cette étude pose les bases pour des recherches futures et des applications pratiques dans le domaine du marketing bancaire, contribuant à l'évolution continue de ce secteur essentiel de l'économie.

Bibliographie

- Allied Market Research, https://www.alliedmarketresearch.com/. (2022). *Data Analytics in Banking Market Size, Share, Competitive Landscape and Trend Analysis Report by Component, by Deployment Model, by Organization Size, by Type, by Application : Global Opportunity Analysis and Industry Forecast, 2021-2031*. Allied Market Research. Récupéré le 14 mai 2024 de https://www.alliedmarketresearch.com/data-analytics-in-banking-market-A16647

- Amazon Web Services. (s.d.). *What is hyperparameter tuning?* AWS. Récupéré le 10 mai 2024 de https://aws.amazon.com/fr/what-is/hyperparameter-tuning/

- Ardelean, A., Burciu, A., & Titan, E. (2013). *The Influence of Quality of Life Indicators on Migration in Europe*. Ovidius University Annals, Economic Sciences Series, 13(1), 2-6. Récupéré de http://stec.univ-ovidius.ro/html/anale/ENG/cuprins%20rezumate/volum2013p1.pdf

- Audzeyeva, A., & Hudson, R. (2016). How to get the most from a business intelligence application during the post implementation phase? Deep structure transformation at a U.K. retail bank. *European Journal Of Information Systems*, 25(1), 29-46. Récupéré de https://doi.org/10.1057/ejis.2014.44

- Boittiaux, P. (2017, 15 mars). L'Europe des services bancaires en ligne. *Statista*. Récupéré le 11 avril 2024 de https://fr.statista.com/infographie/8510/leurope-des-services-bancaires-en-ligne/

- Bourany, T. (2019). Les 5V du big data. *Regards croisés sur l'économie*, n° 23(2), 27-31. Récupéré de https://doi.org/10.3917/rce.023.0027

- Brightwood, S. (2024). *Regularization Techniques in Logistic Regression.* Ladoke Akintola University of Technology. Récupéré le 11 avril 2024 de https://www.researchgate.net/publication/379078786

- Brossault, B. (2023, 1 septembre). Qu'est-ce que le smart data et quelles différences avec le big data ? *Hubspot*. Récupéré le 12 mai 2024 de https://blog.hubspot.fr/marketing/smart-data

- Buzullier, L. (2019). La data dans l'univers bancaire. *Réalités industrielles*, Février 2019(1), 9-13. Récupéré de https://doi.org/10.3917/rindu1.191.0009

- Coret, S. (2021, 27 septembre). *Les dépenses mondiales en matière de big data et d'analyse d'entreprise atteindront 274 milliards de dollars en 2022*. Developpez.com. Récupéré le 12 mai 2024 de https://big-data.developpez.com/actu/318650/Les-depenses-mondiales-en-matiere-de-big-data-et-d-analyse-d-entreprise- atteindront-274-milliards-de-dollars-en-2022-soit-une-hausse-de-27-pourcent-en-un-an-d-apres-Statista-et-IDC/

- Czímer, B., Dietz, M., László, V., & Sengupta, J. (2022). The future of banks : A $ 20 trillion breakup opportunity. *McKinsey & Company*. Récupéré le 20 juin 2024 de https://www.mckinsey.com/industries/financial-services/our-insights/the-future-of-banks-a-20-trillion-dollar-breakup-opportunity

- Data Science and Predictive Analytics – MS in Predictive Analytics & Risk Management. (s. d.). *University of Illinois Board of Trustees*. Récupéré le 17 avril 2024 de https://predictive-analytics.illinois.edu/what-is-data-science-and-what-is-predictive-analytics/

- Dahiru, T. (2011). P-Value, a true test of statistical significance? a cautionary note. *Annals Of Ibadan Postgraduate Medicine*, 6(1). Récupéré de https://doi.org/10.4314/aipm.v6i1.64038

- Davenport, T. (2018). *DELTA Plus Model & Five Stages of Analytics Maturity: A Primer*. International Institute for Analytics. Récupéré le 4 août 2024 de https://iianalytics.com

- Dumoulin, C. (2023, 20 avril). Le big data et ses 3V, un vaste océan d'opportunités. *Business & Decision*. Récupéré le 25 mai 2024 de https://fr.blog.businessdecision.com/3v-opportunites-big-data/

- Erokhin, A. (2022, 2 décembre). Customer2Vec : Representation learning for customer analytics and personalization. Récupéré le 11 mai 2024 de https://www.linkedin.com/pulse/customer2vec-representation-learning-customer-analytics-%D0%B0%D1%80%D1%82%D0%B5%D0%BC-%D0%B5%D1%80%D0%BE%D1%85%D0%B8%D0%BD/

- George, M. (2023). Data Analytics in Marketing - Session 7: Performance & Evaluation. ICHEC. Récupéré le 13 mai 2024 de https://moodle.ichec.be/pluginfile.php/258460/mod_resource/content/0/ICHEC%20-%20Data%20Analytics%20in%20Marketing%20-%20Session%205%20-%2019102023%20-%20Predictive%20Modeling%20I.pdf

- Goetz, É. (2019, 5 août). L'informatique dans la banque, une révolution lente. *Les Echos*. Récupéré le 2 mai 2024 de

https://www.lesechos.fr/2014/08/linformatique-dans-la-banque-une-revolution-lente-1103436

- Gupta, B., Arora, A., Rawat, A., Jain, A., & Dhami, N. (2017). Analysis of Various Decision Tree Algorithms for Classification in Data Mining. *International Journal of Computer Applications*, 163(8). Récupéré le 24 juin 2024 de https://www.ijcaonline.org

- Gupta, S., & Singhal, A. (2019). A Comparative Analysis of Feature Selection Methods for Intrusion Detection System. *International Journal of Computer Applications*, 178(10), 11-16. Récupéré de https://www.ijcaonline.org/archives/volume178/number10/gupta-2019-ijca-918811.pdf

- Great Learning Team. (2024, April 30). *An introduction to GridSearchCV | What is Grid Search.* Great Learning. Récupéré le 24 juin de https://www.mygreatlearning.com/blog/gridsearchcv/

- Hakkarainen, P. (2022, 13 janvier). The digital transformation of the European banking sector: the supervisor's perspective. *European Central Bank - Banking Supervision*. Récupéré le 11 mai 2024 de https://www.bankingsupervision.europa.eu/press/speeches/date/2022/html/ssm.sp220113~8101be7500.en.html

- Hamilton, B. (2018). Transforming Banks into Tech Companies. *Morgan Stanley*. Récupéré le 17 avril 2024 de https://www.morganstanley.com/ideas/banking-digitalization

- Hévin, F. (2023). Marketing bancaire : les coulisses d'un secteur en quête d'innovation et d'adaptation. Récupéré le 17 avril 2024 de https://www.repercom.org/marketing-bancaire-secteur-innovation-adaptation/

- Hotz, N. (2024, April 28). What is CRISP DM? Data Science Process Alliance. Récupéré le 10 juillet 2024 de https://www.datascience-pm.com/crisp-dm-2/

- IBM. (2023). *What is principal component analysis (PCA)?* Récupéré le 9 juillet 2024 de https://www.ibm.com/topics/principal-component-analysis

- Imhoff, C., & White, C. (2011). Self-Service Business Intelligence: Empowering Users to Generate Insights. *TDWI Best Practices Report*, Third Quarter. Récupéré de https://docs.media.bitpipe.com/io_10x/io_106625/item_583281/TDWI_Best_Practices_Report_Self-Service_BI_Q311%5B1%5D.pdf

- Kaur, H., & Geetha, G. (2021). Analysis and Prediction of Different Heart Diseases Using Different Data Mining Techniques. *International Journal of Advanced Trends in Computer*

Science and Engineering, 10(1), 1-7. Récupéré le 23 mai 2024 de http://www.warse.org/IJATCSE/static/pdf/file/ijatcse391012021.pdf

- Kotler, P. and Keller, K.L. (2012) Marketing Management. 14th Edition, *Pearson Education*.

- Kotu, V., & Deshpande, B. (2019). Data Science: Concepts and Practice (2nd ed.). Morgan Kaufmann. Récupéré de https://asolanki.co.in/wp-content/uploads/2019/04/Data-Science-Concepts-and-Practice-2nd-Edition-3.pdf

- Kumari, L., & Aggrawal, D. (2022). An Insight into Predictive Analytics Techniques. *International Journal for Research in Applied Science and Engineering Technology*. Récupéré de https://doi.org/10.22214/ijraset.2022.48071

- Lennerholt, C., Van Laere, J., & Söderström, E. (2018). Implementation Challenges of Self Service Business Intelligence: A Literature Review. *Proceedings Of The Annual Hawaii International Conference On System Sciences* (1999). Récupéré de https://doi.org/10.24251/hicss.2018.631

- Li, Q., Zhao, S., Zhao, S., & Wen, J. (2023). Logistic regression matching pursuit algorithm for text classification. *Knowledge-Based Sys*tems. Récupéré de https://doi.org/10.1016/j.knosys.2023.110761

- Lutz, M., & Lagacherie, M. (2022, 16 septembre). Une histoire de la data science, par deux data scientists. *OCTO Talks !* Récupéré le 1 avril 2024 de https://blog.octo.com/une-histoire-de-la-data-science-par-deux-data-scientists

- Metge, P. (2015). Lebig Dataet La banque. *Revue d'économie financière*, n° 118(2), 93-101. Récupéré de https://doi.org/10.3917/ecofi.118.0093

- Mistrean, L. (2021). CUSTOMER ORIENTATION AS A BASIC PRINCIPLE IN THE CONTEMPORARY ACTIVITY OF THE BANK. *Journal of Public Administration, Finance and Law*. Récupéré de https://doi.org/10.47743/jopafl-2021-21-05.

- Molnar, C. (2023). Interpretable Machine Learning. Récupéré le 18 juin 2024 de https://christophm.github.io/interpretable-ml-book/

- Mousaeirad, S. (2020). Intelligent Vector-based Customer Segmentation in the Banking Industry. *ArXiv*. Récupéré de https://doi.org/10.48550/arXiv.2012.11876

- Muntean, M. (2018). Business Intelligence Issues for Sustainability Projects. *Sustainability*, 10(2), 335. Récupéré de https://doi.org/10.3390/su10020335

- Osei, F., Ampomah, G., Kankam-Kwarteng, C., Bediako, D., & Mensah, R. (2021). Customer Satisfaction Analysis of Banks: The Role of Market Segmentation. *Science Journal of Business and Management*. Récupéré de https://doi.org/10.11648/j.sjbm.20210902.19.

- Osei, L. K., Cherkasova, Y., & Oware, K. M. (2023). Unlocking the full potential of digital transformation in banking: a bibliometric review and emerging trend. *Future Business Journal*, 9(1). Récupéré de https://doi.org/10.1186/s43093-023-00207-2

- Paikra, N. K. (2023). *Bank term deposit subscription dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/neerajkumarpaikra/bank-term-deposit-subscription-dataset?resource=download

- Revolut. (s. d.). Revolut ultra | Revolut Belgique. Récupéré le 6 mai 2024 de https://www.revolut.com/fr-BE/ultra-plan/

- Rozhko, V. I. (2023). Justification of consumer market segmentation as a mandatory tool of strategic marketing. Technology Audit And Production Reserves, 2(4(70)), 15-19. Récupéré de https://doi.org/10.15587/2706-5448.2023.277373

- Sial, A. H., Rashdi, S. Y. S., & Khan, A. H. (2021). Comparative Analysis of Data Visualization Libraries Matplotlib and Seaborn in Python. *International Journal of Advanced Trends in Computer Science and Engineering*, 10(1), 277-281. Récupéré de https://doi.org/10.30534/ijatcse/2021/391012021

- Skender, F. (2023). EVALUATING DATA VISUALIZATION TOOLS BASED ON IMPORTING AND PROCESSING TIMES. *Vision Journal*. Récupéré le 8 juin 2024 de https://visionjournal.edu.mk/social/index.php/1/article/view/141/138

- Sahrir, S. S. (2024). The influence of market segmentation on customer decisions in choosing Bank Syariah Indonesia in North Luwu Regency. Equilibrium, 13(1), 284-291. Récupéré le 28 juin 2024 de https://creativecommons.org/licenses/by-sa/4.0/

- Skender, F., & Manevska, V. (2022). Data Visualization Tools - Preview and Comparison. *Journal of Emerging Computer Technologies*, 2(1), 30-35. Récupéré le 28 mai 2024 de https://www.researchgate.net/publication/362113140_Data_Visualization_Tools_-Preview_and_Comparison

- Thomke, S., & Manzi, J. (2014). The discipline of business experimentation. *Harvard Business Review*. Récupéré le 4 août 2024 de https://hbr.org/2014/12/the-discipline-of-business-experimentation

- Valenti, J, & Alderman, R. (2022, 20 juin). Building on the digital banking momentum. *Deloitte Insights*. Récupéré le 8 mai 2024 de

https://www2.deloitte.com/us/en/insights/industry/financial-services/digitalization-in-banking.html

- Walden, S. (2021, 24 juin). What is a neobank? *Forbes Advisor*. Récupéré le 6 mai 2024 de https://www.forbes.com/advisor/banking/what-is-a-neobank/

- What is Random Forest? | IBM. (s. d.). Récupéré le 20 mai 2024 de https://www.ibm.com/topics/random-forest

- Zaki, A., Khodadadi, N., Lim, W., & Towfek, S. (2024). Predictive Analytics and Machine Learning in Direct Marketing for Anticipating Bank Term Deposit Subscriptions. *American Journal of Business and Operations Research*. Récupéré de https://doi.org/10.54216/ajbor.110110.

- Zhao, S. (2023). Classification and Prediction of Bank Marketing Activity by Machine Learning. *Highlights in Business, Economics and Management*, 21, 725-732. Récupéré de https://doi.org/10.54097/hbem.v21i.14752

- Zu, L., Qi, W., Li, H., Men, X., Lu, Z., Ye, J., & Zhang, L. (2024). UP-SDCG: A Method of Sensitive Data Classification for Collaborative Edge Computing in Financial Cloud Environment. *Future Internet*, 16(3), 102. Récupéré de https://doi.org/10.3390/fi16030102