Projet Personnel

Application Machine Learning : Arbre de décicions

La Désertification Bancaire

Présenté par :
Madou Gagi Ismael

Année universitaire 2024–2025

Table des matières

1 Introduction

La désertion des clients est un problème pour les entreprises de services financiers, y compris les banques. Lorsqu'un client quitte une banque, cela peut avoir un impact négatif sur les revenus de l'entreprise, ainsi que sur sa réputation. Par conséquent, prédire la désertion des clients est essentiel pour les banques afin de maintenir leur rentabilité et leur croissance.

D'après le directeur **banking France** de Diebold Nixdorf(entreprise qui évolue dans le secteur des automates bancaires) connu sous le nom de Grégoire Basquin (2024), avec la montée de la technologie, Effectuer un virement bancaire depuis son smartphone ou échanger avec un conseiller par chat sont devenus pour beaucoup la nouvelle norme concernant leurs besoins bancaires. Cette mutation des comportements bancaires tend à accélérer le phénomène de disparition des agences, dont le nombre a diminué de 14,9 % entre 2009 et 2016. Si les banques ont invoqué comme raison la baisse de fréquentation des agences au profit de la banque en ligne, le manque de rentabilité des agences en est une conséquence directe.

Ce manque de rentabilité des agences pourrait se traduire par une exclusion financière alarmante via la réduction des services comme le retrait au niveau des automates bancaire. Ceci pourrait avoir comme d'énormes conséquences. C'est ce qui nous pousse à nous poser la question suivante : Comment la désertification bancaire affecte-t-elle l'inclusion financière et le développement socio-économique des territoires, et quelles solutions peuvent être mises en œuvre pour garantir un accès équitable aux services bancaires ?

Cette question constitue le socle de notre travail qui sera reparti en trois grandes parties. La première partie sera consacrée au cadre théorique et concept de la désertification bancaire. Puis nous aurons la présentation des données et de la méthodologie de Recherche au niveau de la deuxième partie. Et enfin, les résultats de l'étude seront présentés.

2 La théorie et Concept de la désertification bancaire

2.1 Définition

La désertification bancaire désigne le processus de retrait progressif ou de fermeture des agences bancaires dans certaines zones géographiques, en particulier les territoires ruraux, les petites communes, les quartiers périurbains ou les territoires fragiles sur le plan économique. Il s'agit d'un phénomène lié à la transformation du modèle bancaire, dans un contexte de numérisation accrue, de réduction des coûts et de recherche de rentabilité par les établissements financiers.

Ce phénomène se traduit concrètement par une diminution de l'accès physique aux services bancaires traditionnels (comme les retraits, les dépôts, les conseils personnalisés, la souscription de prêts ou l'ouverture de comptes) au détriment du paiement en ligne. Selon Grégoire Basquin (2024) la disparition des agences a un impact désastreux dans certaines communes qui offrent des services de proximité, facteurs de socialisation optimisant le maillage territorial. Parmi ces services, le retrait d'argent sur automate bancaire reste un appui économique assurant l'accès aux espèces et permettant de favoriser les commerces adjacents. Leur disparition ne ferait qu'accentuer un isolement social et conduire à un déclin démographique. Il affecte en particulier les populations vulnérables, âgées, non connectées ou faiblement bancarisées.

La désertification bancaire soulève plusieurs questions en matière d'inclusion financière, de cohésion territoriale, et de justice sociale, en creusant les inégalités entre les zones bien desservies (souvent urbaines) et les zones en difficulté. Elle contribue à un sentiment d'abandon des services publics et de proximité, et peut freiner le développement économique local.

2.2 Contexte économique et territorial : zones rurales et évolution des services bancaires

2.2.1 Recul des services dans les territoires ruraux

Voilà plusieurs années que les élus assistent à un désengagement bancaire sur les territoires, et en particulier dans les communes rurales. En 2019 déjà, un rapport de la Banque de France montrait que « si les de plus de 5 000 habitants disposent presque toutes d'au moins un distributeur automatique de billets (Dab), la quasi-intégralité de celles de moins de 1 000 habitants n'est pas équipée (0,9 % des communes de moins de 500 habitants en ont au moins un à leur disposition) ». Dans un État des lieux de l'accès du public aux espèces en France métropolitaine, publié le 18 juillet 2022, la Banque de France ne totalise plus que 47 853 automates fin 2021, chiffre en recul de 2 % par rapport à l'année précédente. Les zones rurales connaissent un affaiblissement progressif des services de proximité : fermeture de classes, réduction des horaires de La Poste, raréfaction des médecins, et désormais, disparition des agences bancaires. Ce phénomène s'inscrit dans une dynamique plus large de recomposition des territoires, marquée par :

- **le dépeuplement de certaines communes rurales,**
- **une mobilité contrainte (moins de transports publics),**
- **des difficultés d'accès aux services essentiels.**

Les banques, dans un souci de rationalisation de leurs coûts, concentrent leurs agences dans les zones les plus denses et les plus rentables, au détriment des petites villes ou villages. Cela renforce un sentiment d'inégalité territoriale et d'abandon.

2.2.2 Évolution des services bancaires

Depuis plusieurs années, l'offre bancaire ne cesse d'évoluer. Cette évolution s'inscrit dans une dynamique de transformation numérique, de rationalisation des coûts et de concentration des établissements, entraînant une réduction progressive de la présence physique des agences, notamment dans les zones rurales.

La montée de la technologie a permis aux banques de numériser de plus en plus leur modèle opérationnel Au cours des deux dernières décennies. On observe une dématérialisation massive des services bancaires : Les opérations courantes (virements, paiements, consultation de comptes, édition de RIB, etc.) sont désormais accessibles via Internet ou mobile, 24h/24 et 7j/7 ; Les banques ont investi dans des applications mobiles performantes, avec des fonctionnalités intuitives qui réduisent le besoin de se rendre en agence ; Le selfcare (gestion autonome) devient la norme, réduisant le rôle traditionnel du conseiller bancaire. Cette transition répond à une demande croissante d'autonomie des clients, notamment les plus jeunes, mais elle crée en parallèle une exclusion pour les publics peu familiers avec le numérique (seniors, personnes peu éduquées, habitants de zones peu couvertes par internet).

De plus, comme pour toute entité, l'optimisation des coûts demeure une des préoccupation des banques dans le but d'assurer leurs rentabilités. Ce processus passe par une réduction des agences physiques (ayant un flux de clients limité) qui constituent une charge énorme pour les banques. C'est qui pousse les directions bancaire à réviser leur maillage territorial. Cette baisse du nombre d'agences pourrait être expliquée également par des taux d'intérêts historiquement bas, une forte régulation et une concurrence accrue.

Au même moment, L'émergence des banques en ligne (Boursorama, ING, Hello Bank !, Orange Bank, etc.) et des néobanques (Revolut, N26, Lydia, etc.) a bouleversé le paysage bancaire en ciblant les jeunes génération à la recherche de rapidité et de mobilité. Leur modèle repose sur l'efficacité technologique, la personnalisation automatisée (via l'IA et l'analyse des données clients) et le bas prix. Face à cette situation, les banques traditionnelles se trouvent dans l'obligation d'innover en terme de services proposés aux clients.

Cette modernisation ne s'effectue sans conséquences. Elle a des effets ambivalents : positifs pour pour les territoires urbains, connectés, et dynamiques et négatifs pour les zones rurales.

2.3 Objectifs de la prédiction de la désertification bancaire

La désertification bancaire implique plusieurs enjeux à savoir politiques, économiques et sociaux.

2.3.1 Enjeux politiques

selon PAGES JAUNES (2020), Le rôle principal des banques consiste à assurer la gestion des moyens de paiement. Banque de dépôt, banque d'investissement et de financement, banque privée : il existe plusieurs types d'établissements de crédit. Les banques favorisent le rapprochement entre individus mais facilitent la vie de tous les jours. La disparition de ces banques implique la négligence de la part de l'État ou bien un abandon. Prévoir les zones à risque de désertification permet d'anticiper des inégalités territoriales croissantes et d'adapter les politiques publiques dans le but d'inclure cette partie de la population marginalisée.

2.3.2 Enjeux économiques

Sur le plan économique, la fermeture des agences bancaires entraîne la fermeture des commerces qui va se traduire par une perte d'attractivité et par conséquent un déclin économique local. La prédiction de la désertification bancaire permet aux acteurs économiques et institutionnels de créer un environnement propice au développement des activités comme le commerce,l'agriculture,l'artisanat... Ce processus passe par la distribution bancaire (agences mobiles, distributeurs automatiques, banques mutualistes) et la mise en place des alternatives numériques accompagnées (avec médiateurs ou animateurs digitaux) pour maintenir un tissu économique de proximité.

2.3.3 Enjeux Sociaux

Nous savons que la disparition des agences dans les zones reculées induit une inégalité sociale. Une part importante de la population (notamment les personnes âgées, en situation de précarité, en fracture numérique, etc.) n'utilise pas ou mal les services en ligne.

le manque d'accès aux agences les complique la situation puisqu'ils se verront obligés de se déplacer dans des zones urbaines afin d'accéder aux agences. En anticipant les zones à risque, les pouvoirs publics et les associations peuvent : mettre en place des dispositifs de médiation numérique ; former et accompagner les usagers et créer des guichets polyvalents de proximité pour combler les vides. selon MAIRE INFO (2023), En réponse aux deux questions posées par les sénateurs à propos de la raréfaction des Dab, le gouvernement affirme qu'il lutte contre ce phénomène depuis 2018. Par exemple, Bercy rappelle également dans sa réponse publiée hier que « la loi du 2 juillet 1990 prévoit que La Poste a l'obligation de faire en sorte que, sauf circonstances exceptionnelles, 90 % de la population de chaque département soit éloignée de moins de vingt minutes de trajet automobile, des plus proches points de contact de La Poste. ». Ceci témoigne l'importance de prédire ce phénomène.

2.4 Revue rapide des méthodes de prévision appliquées à la géographie bancaire

La prévision de la désertification bancaire repose sur des outils empruntés à la statistique spatiale, à la modélisation économétrique, et plus récemment à l'intelligence artificielle. Ces méthodes permettent de détecter des zones à risque, d'anticiper les fermetures d'agences, et d'accompagner la prise de décision publique ou privée.

2.4.1 Méthodes statistiques classiques

1. **Régression linéaire / logistique :**

   Selon DATASCIENTEST (2020), La régression logistique est un modèle statistique permettant d'étudier les relations entre un ensemble de variables qualitatives Xi et une variable qualitative Y. Il s'agit d'un modèle linéaire généralisé utilisant une fonction logistique comme fonction de lien. Un modèle de régression logistique permet aussi de prédire la probabilité qu'un événement arrive (valeur de 1) ou non (valeur de 0) à partir de l'optimisation des coefficients de régression. Ce résultat varie toujours entre 0 et 1. Lorsque la valeur prédite est supérieure à un seuil, l'événement est susceptible de se produire, alors que lorsque cette valeur est inférieure au même seuil, il ne l'est pas. Utilisées pour modéliser la probabilité de fermeture d'une agence bancaire en fonction de variables explicatives telles que la densité de population, le revenu moyen, l'usage des services numériques, etc.

   Exemple de modèle de régression logistique :

$$P(\text{Fermeture}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}}$$

2. **Analyse en composantes principales (ACP) :**

   est une méthode de la famille de l'analyse des données qui consiste à transformer des variables liées entre elles (dites « corrélées » en statistique) en nouvelles variables décorrélées les unes des autres (WIKIPÉDIA). Ces nouvelles variables sont nommées « composantes principales » ou axes principaux. Elle permet de résumer l'information en réduisant le nombre de variables. Les champs d'application de l'ACP sont

aujourd'hui multiples, allant de la biologie à la recherche économique et sociale, et plus récemment le traitement d'images et l'apprentissage automatique.

Par exemple On possède des données sur les températures moyennes mensuelles de différentes villes de France et on cherche à résumer le climat français[6]. Intuitivement, les villes les plus chaudes en janvier le sont aussi en juillet. Les variables sont donc corrélées positivement entre elles et l'ACP mettra en évidence un premier axe opposant les villes chaudes aux villes froides. Toutefois, ces variables ne sont pas parfaitement corrélées car il existe des différences entre les villes dans l'écart entre les températures hivernales et estivales. L'ACP dégagera alors un second axe opposant les villes selon l'amplitude des températures : climat océanique ou climat continental.

3. **Cartographie statistique :**

TOBLER, Waldo R dans son article «La cartographie statistique : qu'est-ce que c'est ?», Il existe une longue association historique entre les statistiques et la cartographie, notamment en ce qui concerne la théorie de l'ajustement des observations. La quasi-totalité de cette histoire peut être évoquée en mentionnant simplement le nom de Carl F. Gauss, inventeur de la méthode des moindres carrés. Il existe également une tradition dans laquelle la cartographie prend la forme d'une illustration graphique de données statistiques. Aujourd'hui, on parle souvent de cartographie thématique , parfois de cartographie statistique . En statistique élémentaire, les premières mesures descriptives apprises concernent les tendances centrales. Il y a bien sûr les variantes bidimensionnelles de celles-ci : le centre de gravité, la médiane bivariée, le point de déplacement minimal de l'agrégat ; et les mesures de dispersion (la distance standard, les ellipses bivariées, la cartographie de Mendeleïev et ses extensions par Bachi, et ainsi de suite). Il est même possible d'aller plus loin que ce qui est fait habituellement, jusqu'à la régression bivariée, par exemple en traitant l'image du visage d'un enfant comme une fonction linéaire (ou non linéaire) de l'image du visage de ses parents ou en faisant régresser une carte géographique sur ses précédents historiques. Les outils SIG (systèmes d'information géographique) tels que QGIS permettent de produire des cartes thématiques des zones sous-dotées en services bancaires.

2.4.2 Méthodes économétriques spatiales

Les méthodes économétriques classiques (régression linéaire, logistique...) ne tiennent pas toujours compte de la dimension géographique. Or, dans la désertification bancaire, la proximité spatiale entre territoires influence fortement le phénomène : une commune peut être touchée par la fermeture d'agences simplement parce qu'une zone voisine l'est déjà.

— SAR (Spatial Autoregressive Model) : tient compte de la corrélation entre les zones voisines.

— SEM (Spatial Error Model)

Ici, ce sont les erreurs du modèle qui sont spatialement corrélées. Cela permet de capturer les facteurs géographiques non observés (ex. : politiques locales).

— GWR (Geographically Weighted Regression) :

Contrairement à la régression classique, GWR adapte les coefficients des variables explicatives localement : les déterminants de la désertification peuvent varier d'un territoire à l'autre. Exemple : dans une commune rurale, l'âge moyen peut jouer un rôle fort, alors que dans une commune périurbaine, c'est plutôt l'usage du numérique. Ces modèles permettent de produire des cartes prédictives précises de la désertification bancaire.

2.4.3 Méthodes d'apprentissage automatique (machine learning)

De plus en plus utilisées pour identifier les zones à risque à partir de grands volumes de données hétérogènes. Les méthodes de machine learning sont particulièrement adaptées pour prédire un phénomène complexe et multifactoriel comme la désertification bancaire, à partir de données massives et hétérogènes.

1. **Arbres de décision & forêts aléatoires (Random Forest)** : Techniques supervisées de classification ou de régression. Cette méthode est utilisée pour prédire dans notre cas si une commune est à risque de désertification bancaire à partir de plusieurs indicateurs (âge moyen, revenu, couverture Internet, etc.).
2. **XGBoost** : XGBoost (Extreme Gradient Boosting) est un algorithme d'apprentissage supervisé basé sur le principe du boosting par gradient, qui consiste à entraîner successivement plusieurs arbres de décision faibles, chacun corrigeant les erreurs du précédent. Très utilisé en data science pour sa précision, sa vitesse d'exécution et sa capacité à éviter le surapprentissage grâce à des techniques de régularisation (L1, L2), XGBoost est particulièrement performant sur les données tabulaires. Il prend en charge les valeurs manquantes, offre une gestion automatique de la complexité des modèles, et s'intègre facilement avec des bibliothèques comme scikit-learn. Couramment utilisé dans des domaines comme la finance, la santé ou le marketing, XGBoost permet par exemple de prédire la probabilité de désertion bancaire à partir de variables comme l'âge, l'usage numérique ou le revenu du client.
3. **K-means (Clustering)** : K-means est un algorithme de clustering non supervisé qui permet de regrouper des données en K groupes homogènes selon leur similarité. Il fonctionne en assignant chaque point de données au centre (centroïde) le plus proche, puis en recalculant ces centres de manière itérative jusqu'à stabilisation. L'objectif est de minimiser la variance intra-groupe (somme des distances entre les points et leur centroïde). Simple, rapide et efficace sur les données volumineuses, K-means est largement utilisé pour segmenter des populations (ex. : clients, zones géographiques), détecter des profils types ou analyser des comportements. En contexte bancaire, il peut servir à regrouper les agences ou les communes selon des critères socio-économiques pour identifier des zones à risque de désertification, ou pour adapter l'offre de services à chaque segment identifié.

2.4.4 Méthodes hybrides et prospectives

Certaines recherches adoptent des approches hybrides et prospectives. Celles-ci combinent méthodes quantitatives et données qualitatives issues du terrain (enquêtes auprès des habitants, entretiens avec les élus locaux, etc.) afin de mieux comprendre les dynamiques à l'œuvre. On trouve également des méthodes de scénarisation prospective, qui visent à construire différents futurs possibles en fonction de l'évolution des technologies, des politiques publiques ou des besoins sociaux. Cela permet par exemple d'imaginer l'impact du développement des services bancaires mobiles ou des maisons France Services

dans les territoires isolés. L'usage de cartes de risques ou d'indicateurs composites permet de visualiser les zones les plus menacées par la désertification et de cibler les actions de prévention ou de compensation. Ces outils sont essentiels pour anticiper les inégalités territoriales croissantes et adapter les politiques d'aménagement du territoire.

2.5 Présentation de l'approche choisie : arbres de décision

Les méthodes arborescentes sont simples et utiles pour l'interprétation. Cependant, elles ne sont généralement pas compétitives par rapport aux meilleures approches d'apprentissage supervisé. Selon JAMES, Gareth, WITTEN, Daniela, HASTIE, Trevor et TIBSHIRANI, Robert (2013), Les arbres de décision peuvent être appliqués aux problèmes de régression et de classification. Nous examinerons d'abord les problèmes de régression, puis nous passerons à la classification.

A. Arbre de Regression

2.5.1 Prédiction via la stratification de l'espace des caractéristiques

Selon JAMES, Gareth, WITTEN, Daniela, HASTIE, Trevor et TIBSHIRANI, Robert (2013), Nous avons deux étapes pour créer l'arbre de décisions à savoir :

— Diviser l'espace des prédicteurs, c'est-à-dire l'ensemble des valeurs possibles pour X1, X2, . . . , Xp, en J régions distinctes et non chevauchantes, R1, R2, . . . , RJ.

— Pour chaque observation comprise dans la région Rj, nous faisons la même prédiction, qui est simplement la moyenne des valeurs de réponse pour les observations d'apprentissage dans Rj.

En théorie, les régions pourraient avoir n'importe quelle forme. Cependant pour construire les régions, nous choisissons de diviser l'espace prédictif en rectangles de grande dimension, ou boîtes, pour simplifier et faciliter l'interprétation du modèle prédictif obtenu. L'objectif est de trouver les boîtes R1, . . . , RJ qui minimisent le RSS, donné par :

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \tag{1}$$

ù ˆ yRj est la réponse moyenne des observations d'apprentissage dans la j-ième boîte. Malheureusement, il est informatiquement impossible d'envisager toutes les partitions possibles de l'espace des caractéristiques en J boîtes. C'est pourquoi nous adoptons une approche descendante et gourmande, appelée découpage binaire récursif. Cette approche est descendante car elle commence au sommet de l'arbre (où toutes les observations appartiennent à une même région) et découpe successivement l'espace des prédicteurs ; chaque division est indiquée par deux nouvelles branches plus bas dans l'arbre. Elle est gourmande car, à chaque étape du processus de construction de l'arbre, la meilleure division est effectuée à cette étape particulière, plutôt que de se projeter dans l'avenir et de choisir une division qui conduira à un meilleur arbre à une étape ultérieure.

Pour effectuer une décomposition binaire récursive, nous sélectionnons d'abord le prédicteur Xj et le point de coupure s tels que la décomposition de l'espace des prédicteurs en régions

$$\{X \mid X_j < s\} \quad \text{et} \quad \{X \mid X_j \geq s\}$$

entraîne la plus grande réduction possible du RSS. (La notation X|Xj < s désigne la région de l'espace des prédicteurs où Xj prend une valeur inférieure à s.) Autrement dit, nous considérons tous les prédicteurs X1, . . . , Xp et toutes les valeurs possibles du point de coupure s pour chacun des prédicteurs, puis nous choisissons le prédicteur et le point de coupure tels que l'arbre résultant présente le RSS le plus faible. Plus précisément, pour tout j et s, nous définissons la paire de demi-plans

$$R\_1(j, s) = \{X \mid X\_j < s\}, \quad R\_2(j, s) = \{X \mid X\_j \geq s\} \tag{2}$$

Et on cherche les valeurs de j et s qui minimisent l'expression suivante :

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 \tag{3}$$

où ˆ yR1 est la réponse moyenne des observations d'apprentissage dans R1(j, s), et ˆ yR2 est la réponse moyenne des observations d'apprentissage dans R2(j, s). Trouver les valeurs de j et s qui minimisent (3) est assez rapide, surtout lorsque le nombre de caractéristiques p n'est pas trop élevé. Nous répétons ensuite le processus en recherchant le meilleur prédicteur et le meilleur point de coupure afin de fractionner davantage les données et de minimiser le RSS dans chacune des régions obtenues. Cependant, cette fois, au lieu de fractionner l'espace des prédicteurs dans son intégralité, nous fractionnons l'une des deux régions précédemment identifiées. Nous avons maintenant trois régions. Nous cherchons à nouveau à fractionner davantage l'une de ces trois régions afin de minimiser le RSS. Le processus se poursuit jusqu'à ce qu'un critère d'arrêt soit atteint ; par exemple, nous pouvons continuer jusqu'à ce qu'aucune région ne contienne plus de cinq observations. Une fois les régions R1,. . . , RJ ayant été créés, nous prédisons la réponse pour une observation de test donnée en utilisant la moyenne des observations d'entraînement dans la région à laquelle appartient cette observation de test.

2.5.2 Elagage des arbres

JAMES, Gareth, WITTEN, Daniela, HASTIE, Trevor et TIBSHIRANI, Robert (2013) stipulent que le processus décrit ci-dessus peut produire de bonnes prédictions sur l'ensemble d'entraînement, mais risque de sur-ajuster les données, ce qui entraîne de mauvaises performances de l'ensemble de test. Cela est dû au fait que l'arbre résultant pourrait être trop complexe. Un arbre plus petit, avec moins de divisions (c'est-à-dire moins de régions R1, . . . , RJ), pourrait entraîner une variance plus faible et une meilleure interprétation, au prix d'un léger biais. Une alternative possible au processus décrit ci-dessus consiste à construire l'arbre uniquement tant que la diminution du RSS due à chaque division dépasse un seuil (élevé). Cette stratégie produira des arbres plus petits, mais elle est trop imprévoyante, car une division apparemment inutile au début de l'arbre pourrait être suivie d'une très bonne division, c'est-à-dire une division entraînant une forte réduction du RSS ultérieurement.

Par conséquent, une meilleure stratégie consiste à développer un très grand arbre T0, puis à l'élaguer pour obtenir un sous-arbre. Comment déterminer la meilleure méthode d'élagage ? Intuitivement, notre objectif est de sélectionner le sous-arbre qui génère le taux d'erreur de test le plus faible. Étant donné un sous-arbre, nous pouvons estimer son erreur de test par validation croisée ou par l'approche de l'ensemble de validation. Cependant, estimer l'erreur de validation croisée pour chaque sous-arbre possible serait trop complexe, car le nombre de sous-arbres possibles est extrêmement élevé. Nous avons

donc besoin d'un moyen de sélectionner un petit ensemble de sous-arbres à considérer. L'élagage par complexité de coût, également appelé élagage du maillon le plus faible, nous permet d'y parvenir. Plutôt que de considérer chaque sous-arbre possible, nous considérons une séquence d'arbres indexée par un paramètre de réglageαnon négatif.

Algorithme : Construire un arbre de régression

1. Utiliser la division binaire récursive pour développer un grand arbre à partir des données d'apprentissage, en s'arrêtant uniquement lorsque chaque nœud terminal présente un nombre d'observations inférieur à un certain nombre minimal.
2. Appliquer un élagage de complexité de coût au grand arbre afin d'obtenir une séquence de meilleurs sous-arbres, en fonction de α.
3. Utiliser la validation croisée K-fold pour choisir α. Autrement dit, diviser les observations d'apprentissage en K-folds. Pour chaque $k = 1, \ldots, K$:
   (a) Répéter les étapes 1 et 2 sur toutes les données d'apprentissage sauf le k-ième fold.
   (b) Évaluer l'erreur quadratique moyenne de prédiction sur les données du k-ième fold non sélectionné, en fonction de α.

   Faire la moyenne des résultats pour chaque valeur de α et choisir α pour minimiser l'erreur moyenne.
4. Renvoyer le sous-arbre de l'étape 2 correspondant à la valeur choisie de α.

À chaque valeur de α correspond un sous-arbre $T \subset T_0$, obtenu par élagage de coût-complexité.

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T| \tag{4}$$

soit le plus petit possible. Ici, $|T|$ indique le nombre de nœuds terminaux de l'arbre T, Rm est le rectangle (c'est-à-dire le sous-ensemble de l'espace des prédicteurs) correspondant au m-ième nœud terminal, et ˆ yRm est la réponse prédite associée à Rm, c'est-à-dire la moyenne des observations d'apprentissage dans Rm. Le paramètre de réglage α contrôle un compromis entre la complexité du sous-arbre et son ajustement aux données d'apprentissage. Lorsque $\alpha = 0$, le sous-arbre T est simplement égal à T0, car (4) mesure alors simplement l'erreur d'apprentissage. Cependant, à mesure que α augmente, un arbre comportant de nombreux nœuds terminaux a un prix : la quantité (4) tend donc à être minimisée pour un sous-arbre plus petit. Il s'avère que lorsque α augmente à partir de zéro dans (4), les branches sont élaguées de l'arbre de manière imbriquée et prévisible, ce qui facilite l'obtention de la séquence complète de sous-arbres en fonction de α. Nous pouvons sélectionner une valeur de α à l'aide d'un ensemble de validation ou d'une validation croisée. Nous revenons ensuite à l'ensemble de données complet et obtenons le sous-arbre correspondant à α. Ce processus est résumé dans l'algorithme.

Notre projet repose sur la prédiction de la desertification bancaire, par consequent, nous nous focaliserons que sur la prédiction avec l'arbre de décision. Cependant, Retenez qu'il est possible d'effectuer la classification avec cette méthode.

B. Classification avec l'arbre de décision

JAMES, Gareth, WITTEN, Daniela, HASTIE, Trevor et TIBSHIRANI, Robert (2013), Un arbre de classification est très similaire à un arbre de régression, à la différence près qu'il est utilisé pour prédire une réponse qualitative plutôt que quantitative. Rappelons que pour un arbre de régression, la réponse prédite pour une observation est donnée par la réponse moyenne des observations d'apprentissage appartenant au même nœud terminal. En revanche, pour un arbre de classification, nous prédisons que chaque observation appartient à la classe d'observations d'apprentissage la plus fréquente dans la région à laquelle elle appartient. Lors de l'interprétation des résultats d'un arbre de classification, nous nous intéressons souvent non seulement à la prédiction de classe correspondant à une région de nœud terminal particulière, mais aussi aux proportions de classe parmi les observations d'apprentissage qui appartiennent à cette région. Développer un arbre de classification est assez similaire à développer un arbre de régression. Comme dans le cas de la régression, nous utilisons la division binaire récursive pour développer un arbre de classification. Cependant, dans le cas de la classification, le RSS ne peut pas être utilisé comme critère pour effectuer les divisions binaires. Une alternative naturelle au RSS est le taux d'erreur de classification. Puisque nous planifions la classification pour attribuer une observation d'une région donnée à la classe d'observations d'apprentissage la plus courante dans cette région, le taux d'erreur de classification est simplement la fraction des observations d'apprentissage de cette région qui n'appartiennent pas à la classe la plus courante :

$$E = 1 - \max_k (\hat{p}_{mk}) \tag{5}$$

Ici, $\hat{p}_{mk}$ représente la proportion d'observations d'apprentissage de la m-ième région appartenant à la k-ième classe. Cependant, il s'avère que l'erreur de classification n'est pas suffisamment sensible pour la croissance des arbres, et en pratique, deux autres mesures sont préférables. L'indice de Gini est défini par :

$$G = \sum_{k=1}^{K} \hat{p}_{mk} (1 - \hat{p}_{mk}) \tag{6}$$

une mesure de la variance totale entre les K classes. Il est facile de constater que l'indice de Gini prend une faible valeur si tous les $\hat{p}_{mk}$ sont proches de zéro ou de un. C'est pourquoi l'indice de Gini est considéré comme une mesure de la pureté des nœuds : une faible valeur indique qu'un nœud contient principalement des observations d'une seule classe.

Une alternative à l'indice de Gini est l'entropie, donnée par l'entropie :

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk} \tag{7}$$

Puisque $0 \leq \hat{p}_{mk} \leq 1$, il en resulte que $0 \leq -\hat{p}_{mk} \log \hat{p}_{mk}$. On peut montrer que l'entropie prendra une valeur proche de zéro si les p^mk sont tous proches de zéro ou de 1. Par conséquent, comme l'indice de Gini, l'entropie prendra une faible valeur si le m-ième nœud est pur. En fait, il s'avère que l'indice de Gini et l'entropie sont numériquement assez similaires. En construisant un arbre de classification, soit l'indice de Gini ou L'entropie est généralement utilisée pour évaluer la qualité d'une division particulière, car ces deux

approches sont plus sensibles à la pureté des nœuds que le taux d'erreur de classification. Chacune de ces trois approches peut être utilisée lors de l'élagage de l'arbre, mais le taux d'erreur de classification est préférable si l'objectif est la précision de prédiction de l'arbre final élagué.

C. Arbres contre modèles linéaires

Les arbres de régression et de classification ont une nature très différente des approches plus classiques de régression et de classification. En particulier, la régression linéaire suppose un modèle de la forme :

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{8}$$

alors que les arbres de régression supposent un modèle de la forme :

$$f(X) = \sum_{m=1}^{M} c_m \cdot \mathbf{1}(X \in R_m) \tag{9}$$

où R1, . . . , RM représentent une partition de l'espace des caractéristiques. Quel modèle est le meilleur ? Cela dépend du problème traité. Si la relation entre les caractéristiques et la réponse est bien approximée par un modèle linéaire , alors une approche telle que la régression linéaire sera probablement efficace et surpassera une méthode telle qu'un arbre de régression qui n'exploite pas cette structure linéaire. En revanche, s'il existe une relation hautement non linéaire et complexe entre les caractéristiques et la réponse, alors les arbres de décision peuvent surpasser les approches classiques. Les performances relatives des approches arborescentes et classiques peuvent être évaluées en estimant l'erreur de test, soit par validation croisée, soit par l'approche par ensemble de validation. Bien entendu, d'autres considérations que la simple erreur de test peuvent entrer en jeu dans le choix d'une méthode d'apprentissage statistique ; par exemple, dans certains contextes, la prédiction à l'aide d'un arbre peut être préférée pour des raisons d'interprétabilité et de visualisation.

D. Avantages et inconvénients des arbres

Les arbres de décision pour la régression et la classification présentent de nombreux avantages par rapport aux approches plus classiques.

— Les arbres sont très faciles à expliquer. En fait, ils sont même plus faciles à expliquer que la régression linéaire !

— Certains pensent que les arbres de décision reflètent davantage la prise de décision humaine que les approches classiques de régression et de classification.

— Les arbres peuvent être représentés graphiquement et sont facilement interprétés même par un non-expert (surtout s'ils sont petits).

— Les arbres peuvent facilement gérer des prédicteurs qualitatifs sans avoir besoin de créer des variables fictives.

Ils presentent aussi des inconvenients.

— Malheureusement, les arbres n'ont généralement pas le même niveau de précision prédictive que certaines autres approches de régression et de classification présentées dans ce livre.

— De plus, les arbres peuvent être très peu robustes. Autrement dit, une légère modification des données peut entraîner une modification importante de l'arbre final estimé.

3 Présentation de la base de données

La base de données utilisé dans ce projet a été obtenu à partir du site https ://leaps.analyttica.com/ho
Il s'agit d'un ensemble de données sur les clients d'une banque, contenant des informations telles que l'âge, le salaire, l'état matrimonial, la limite de la carte de crédit, la catégorie de carte de crédit, etc. Il y a presque 18 fonctionnalités au total. Pour notre étude nous allons garder 20 variables.

Le dataset comprend 10 000 clients, dont seulement 16,07% ont quitté la banque, ce qui rend la prédiction du churn des clients plus difficile en raison du déséquilibre de classe.

3.1 Description des variables

Les variables sont regroupées en plusieurs catégories : démographiques, financières, comportementales, etc.

`CLIENTNUM`
: Identifiant unique du client, noté $\mathrm{ID}_{\text{client}} \in \mathbb{N}$.

`Attrition_Flag`
: Variable cible, indiquant si le client a quitté la banque :

$$\text{Attrition\_Flag} = \begin{cases} 1 & \text{si le client a quitté la banque (attrition)} \\ 0 & \text{si le client est resté} \end{cases}$$

`Customer_Age`
: Âge du client, Age $\in \mathbb{N}$, exprimé en années.

`Gender`
: Sexe du client, Gender $\in \{\text{Male}, \text{Female}\}$.

`Dependent_count`
: Nombre de personnes à charge (enfants, conjoint...), noté $D \in \mathbb{N}$.

`Education_Level`
: Niveau d'éducation du client, $E \in \{\text{High School}, \text{Graduate}, \text{Doctorate}, \ldots\}$.

`Marital_Status`
: Statut matrimonial, $M \in \{\text{Single}, \text{Married}, \text{Divorced}, \ldots\}$.

`Income_Category`
: Tranche de revenu annuel du client, notée $R \in \{< \$40\text{K}, \$40\text{K}–\$60\text{K}, \ldots\}$.

`Card_Category`
: Type de carte bancaire détenue, $C \in \{\text{Blue}, \text{Silver}, \text{Gold}, \text{Platinum}\}$.

`Months_on_book`
: Ancienneté du compte en mois, soit $MOB \in \mathbb{N}$.

`Months_Inactive_12_mon`
: Nombre de mois d'inactivité dans les 12 derniers mois, $I_{12} \in [0, 12]$.

`Contacts_Count_12_mon`
: Nombre de contacts client-service sur les 12 derniers mois, $CC_{12} \in \mathbb{N}$.

`Credit_Limit`
: Limite de crédit autorisée, exprimée en dollars, $L \in \mathbb{R}^+$.

`Total_Revolving_Bal`
: Solde renouvelable total (non remboursé en fin de période), noté $RB \in \mathbb{R}^+$.

`Avg_Open_To_Buy`
: Moyenne de la réserve disponible = limite de crédit moins le solde renouvelable :

$$\text{OTB} = L - RB$$

`Total_Amt_Chng_Q4_Q1`
: Variation relative du montant total des transactions entre le 4$^{\text{e}}$ et le 1$^{\text{er}}$ trimestre :

$$\Delta_{\text{amt}} = \frac{T_4 - T_1}{T_1}$$

`Total_Trans_Amt`
: Montant total des transactions sur la période, $T_{\text{amt}} \in \mathbb{R}^+$.

`Total_Trans_Ct`
: Nombre total de transactions effectuées, $T_{\text{ct}} \in \mathbb{N}$.

`Total_Ct_Chng_Q4_Q1`
: Variation relative du nombre de transactions entre T4 et T1 :

$$\Delta_{\text{ct}} = \frac{Ct_4 - Ct_1}{Ct_1}$$

`Avg_Utilization_Ratio`
: Ratio moyen d'utilisation du crédit :

$$\text{Ratio}_{\text{utilisation}} = \frac{RB}{L}$$

3.2 Analyse Univariée

A ce niveau, nous explorons les différentes statistiques afin de bien sélectionner les variables à modéliser.

3.2.1 Variable Cible

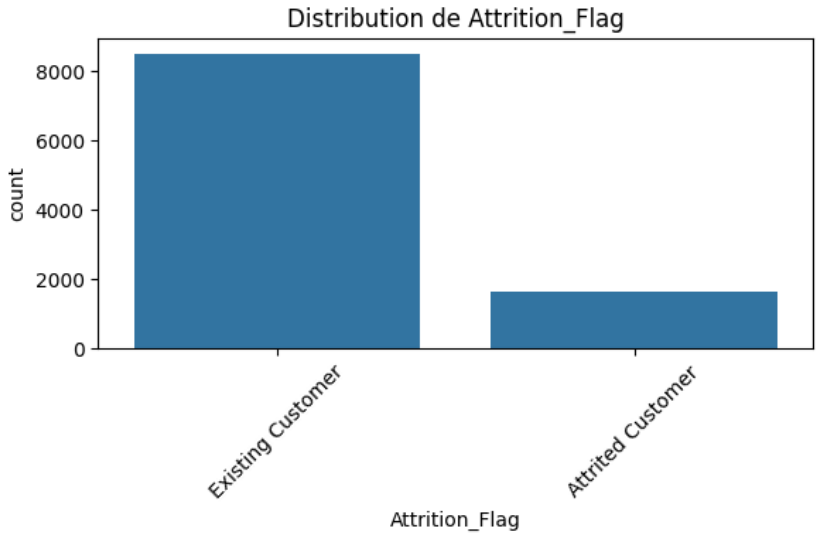

FIGURE 1 – Histogramme des valeurs de la variable cible

Il s'agit de la variable cible scindée en deux classes. Cependant, nous observons un certain déséquilibre de classe où une classe de la variable cible est nettement plus fréquente que l'autre. La classe majoritaire dans cet ensemble de données est "Existing Customer", ce qui signifie qu'il y a un grand nombre d'observations de clients existants par rapport aux clients non-existants. Ce déséquilibre de classe peut poser des problèmes dans certaines tâches d'apprentissage automatique, car les modèles peuvent avoir tendance à être biaisés en faveur de la classe majoritaire, ce qui peut entraîner une performance médiocre pour l'autre classe. Pour cela, nous allons procéder à une des techniques de rééquilibrage comme le sous-échantillonnage (undersampling), sur-échantillonnage (oversampling) ou encore SMOTE (Synthetic Minority Over-sampling Technique). Dans notre cas, nous allons utiliser undersampling pour réequilibrer les classes.

3.2.2 Variable : niveau de l'éducation

(Fig 2)

Nous remarquons que Graduate et High School sont les plus fréquents suivis par Uneducated et Unknown(Sans éducation) ont des effectifs similaires, se situant tous deux entre 500 et 1000 personnes. Il est notable que le groupe "Sans éducation" constitue encore un segment substantiel. Les détenteurs de diplômes avancés "Post-Graduate" (Post-universitaire) et "Doctorate" (Doctorat) sont moins nombreux, ce qui est attendu dans une population générale. Leurs effectifs sont inférieurs à 500. La variable Unknown represente des valeurs manquantes. C'est un facteur important pour toute analyse, car la manière de traiter ces valeurs manquantes (imputation, suppression...) pourrait avoir un impact significatif sur les résultats.

3.2.3 Variable : Genre

Le graphique (Fig 3) est un diagramme à barres (ou diagramme de comptage) qui visualise la répartition des individus selon leur genre. Contrairement à la distribution de l'éducation qui montrait des groupes majoritaires, la répartition par genre ici est équilibrée. Le nombre d'hommes ("Male") et de femmes ("Female") est pratiquement identique, avec une différence infime, peut-être de l'ordre de quelques dizaines d'individus sur un total de plusieurs milliers. Cette distribution équilibrée suggère que le jeu de données n'est pas biaisé vers un genre en particulier. Pour de nombreuses analyses, cela signifie que les tendances observées ne seront pas faussées par une surreprésentation d'un groupe.

3.2.4 Variable : Catégorie des revenus

Le graphique (fig 4) illustre la répartition des clients selon différentes catégories de revenus annuels. La catégorie de revenu "Less than $40K" (Moins de 40 000 $) est de loin la plus importante, représentant le plus grand nombre d'individus dans l'échantillon (environ 3500). Cela suggère que la base de clients est majoritairement composée de personnes à revenus modestes. Les tranches "$40K - $60K" et "$60K - $80K" possèdent une taille similaire et significative, formant la classe moyenne de cette clientèle. La catégorie "$120K +" reste la plus faible. La catégorie "Unknown" (Inconnu) rivalise presque en taille avec le groupe "$40K - $60K". Un volume aussi élevé de données manquantes peut fausser toute analyse et doit être traité avec soin.

3.2.5 Variable : Catégorie des Cartes

Ce diagramme (fig 5) nous renseigne sur la répartition des catégories en fonction de la nature de la carte. La catégorie "Blue" (Bleue) est ultra-majoritaire. Elle représente la vaste majorité de la base clients, avec un effectif estimé à plus de 9000 individus, éclipsant complètement toutes les autres catégories. Les cartes "Silver" (Argent) et "Gold" (Or) sont beaucoup plus rares, représentant chacune une petite fraction de la base. La carte "Platinum" (Platine) est la plus rare de toutes, ce qui est cohérent avec le fait qu'il s'agit généralement de la carte la plus exclusive et exigeante en termes de revenus ou de dépenses. Ce portefeuille est typique d'une banque où la carte "Blue" est le produit d'appel, probablement proposée par défaut ou avec des critiques d'éligibilité faibles. L'objectif commercial principal est très probablement de faire monter en gamme ("up-selling") les détenteurs de carte Blue vers des cartes supérieures plus rémunératrices.

3.2.6 Variable : statut matrimonial

Le graphique (fig 6) visualise la répartition des clients selon leur statut matrimonial. La catégorie "Married" (Marié) est de loin la plus importante, représentant le plus grand nombre d'individus dans l'échantillon. Cela indique que la majorité de la clientèle déclare être mariée. La catégorie "Single" (Célibataire) constitue le deuxième groupe le plus important, formant une part substantielle de la base clients. Les catégories "Divorced" (Divorcé) et "Unknown" (Inconnu) sont significativement plus petites que les groupes "Married" et "Single". La présence d'une catégorie "Unknown" indique que le statut matrimonial n'a pas été renseigné pour une partie des clients. Bien que ce groupe soit petit, sa présence doit être notée pour évaluer la qualité des données.

3.3 Variable : Variables numériques

L'analyse des histogrammes (fig 7) montre que les clients sont majoritairement âgés de 40 à 50 ans, avec une ancienneté moyenne d'environ trois ans, et ont en général entre un et trois dépendants. La plupart présentent deux à trois mois d'inactivité annuelle et effectuent deux à trois contacts avec la banque. Les limites de crédit et l'« open to buy » sont fortement asymétriques, concentrées sur de faibles valeurs avec quelques clients bénéficiant de montants très élevés. Les soldes révolvings et les montants de transactions révèlent une distribution multimodale, traduisant l'existence de segments de clients aux comportements différents. Les changements entre trimestres (montants et transactions) sont globalement faibles et concentrés, tandis que le ratio d'utilisation du crédit reste bas pour la majorité. Ces résultats suggèrent une population relativement homogène sur certains aspects (âge, contacts, inactivité), mais hétérogène en termes d'usage du crédit et de volume de transactions, ce qui pourrait être déterminant dans l'analyse du churn.

4 Bibliographie

LA GAZETTE DES COMMUNES, 2024. *La désertification bancaire : une tendance qui doit s'inverser*. [en ligne]. Disponible sur : https://www.lagazettedescommunes.com/654938/la-desertification-bancaire-une-tendance-qui-doit-sinverser [Consulté le 25 juillet 2025].

MAIRE INFO. *La désertification bancaire menace les territoires ruraux* [en ligne]. 23 juin 2023. Disponible à l'adresse : https://www.maire-info.com/la-desertification-bancaire-\menace-les-territoires-ruraux-article2-26926 (consulté le 26 juillet 2025).

PAGES JAUNES. *Banque : définition* [en ligne]. 20 juillet 2020. Disponible à l'adresse : https://banque.pagesjaunes.fr/comprendre/banque-definition (consulté le 26 juillet 2025).

WIKIPÉDIA. *Analyse en composantes principales* [en ligne]. Disponible à l'adresse : https://fr.wikipedia.org/wiki/Analyse_en_composantes_principales (consulté le 26 juillet 2025).

DATASCIENTEST. *La régression logistique, qu'est-ce que c'est ?* [en ligne]. 4 novembre 2020. Disponible à l'adresse : https://datascientest.com/regression-logistique-quest-ce-que- (consulté le 26 juillet 2025).

TOBLER, Waldo R. *La cartographie statistique : qu'est-ce que c'est ?* [en ligne]. Santa Barbara : University of California. HAL, hal-03739509v1. Disponible à l'adresse : https://hal.science/hal-03739509v1/file/doc00034946.pdf (consulté le 26 juillet 2025).

JAMES, Gareth, WITTEN, Daniela, HASTIE, Trevor et TIBSHIRANI, Robert. *An Introduction to Statistical Learning : with Applications in R*. New York : Springer, 2013. (Springer Texts in Statistics). DOI : https://doi.org/10.1007/978-1-4614-7138-7.

5 Annexe

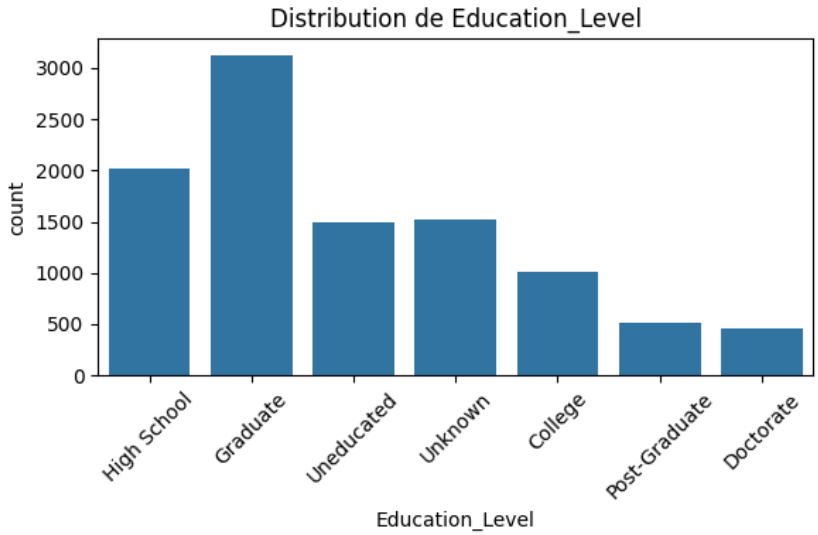

FIGURE 2 – Histogramme des valeurs de la variable éducation

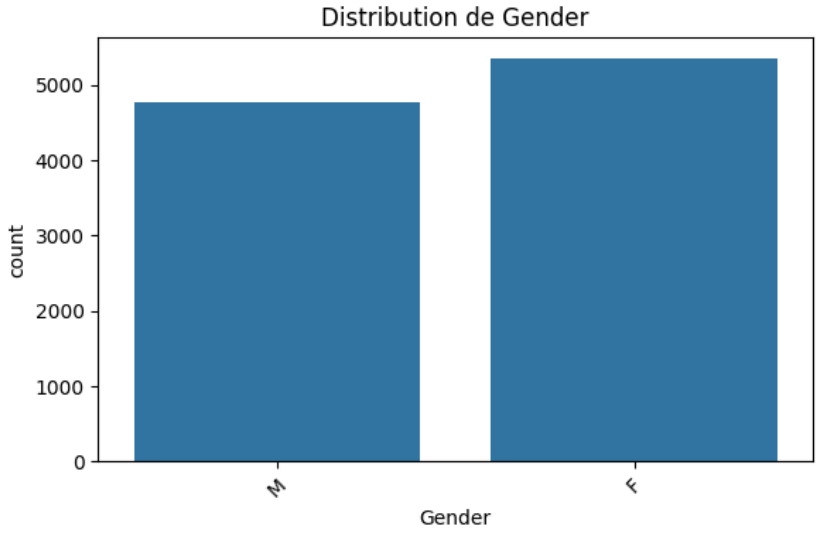

FIGURE 3 – Histogramme des valeurs de la variable Genre

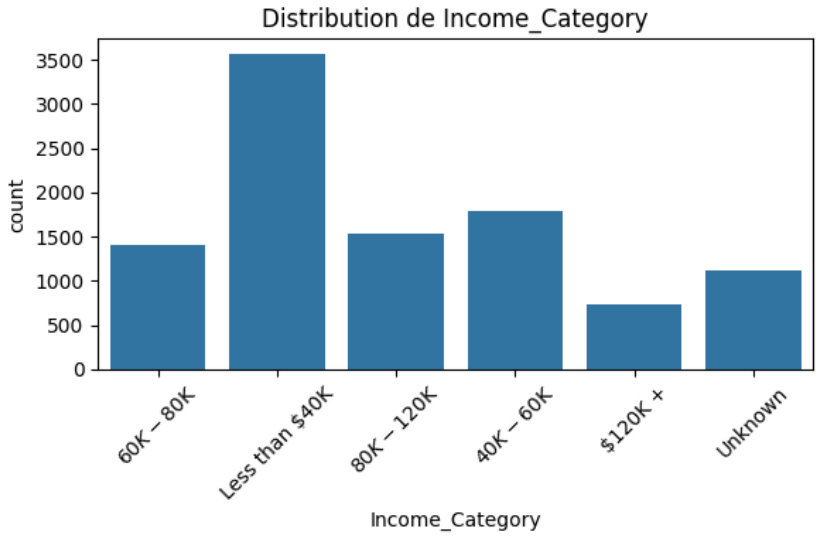

FIGURE 4 – Histogramme des valeurs de la variable Revenu

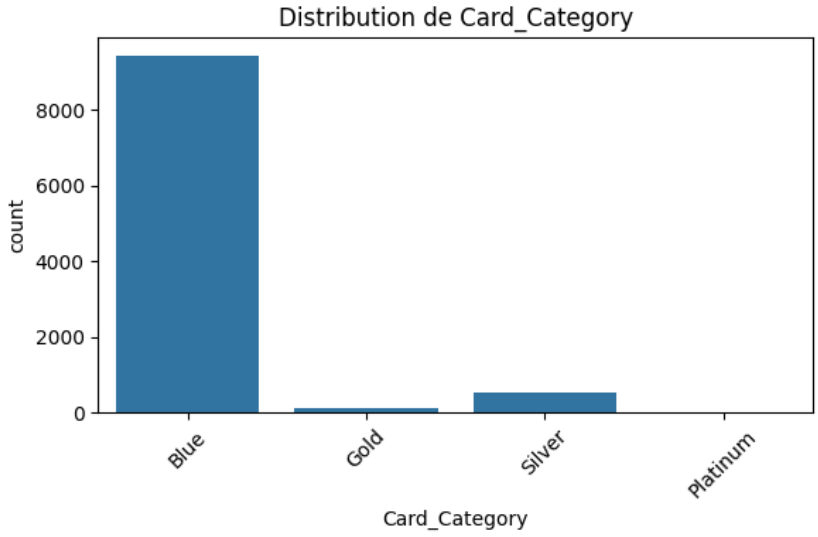

FIGURE 5 – Histogramme des valeurs de la variable Type de carte bancaire

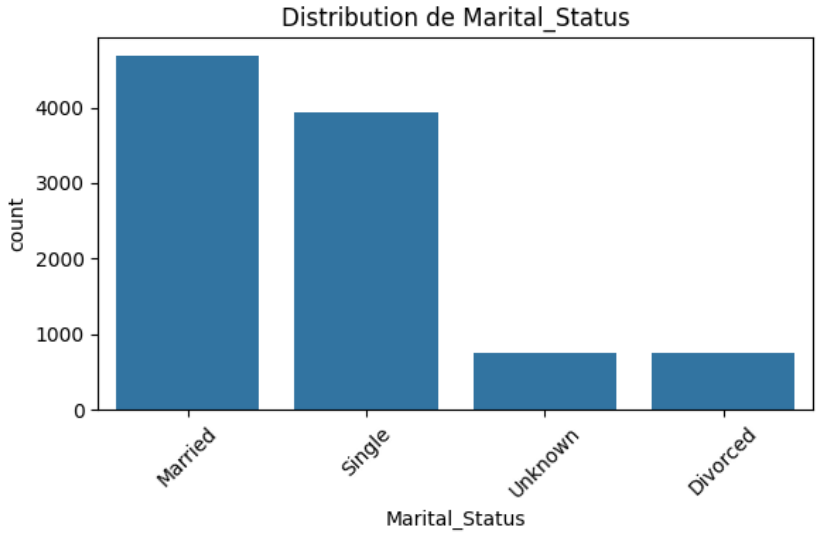

FIGURE 6 – Histogramme des valeurs de la variable Statut matrimonial

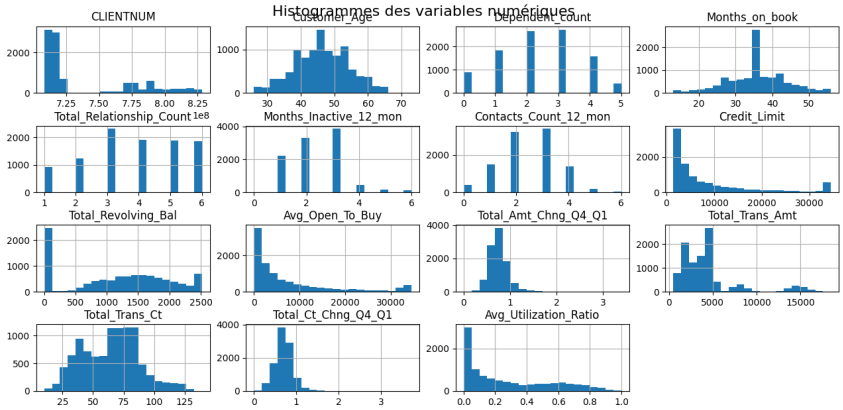

FIGURE 7 – Histogramme des valeurs des variables numériques