Remerciements

Je tiens à exprimer ma plus profonde gratitude à ma mentor et superviseure, Solène Bienaise Biesok, responsable de l'équipe EDA chez BNP Paribas. Son expertise en tant que statisticienne et data scientist, ainsi que son encadrement et son soutien, ont été inestimables tout au long de mon stage.

Je souhaite également remercier mes collègues, Roxane Douc, Abidjo, et Arthur, pour leur assistance avec les questions techniques et commerciales. Votre volonté de partager vos connaissances et vos idées a été grandement appréciée.

Un merci tout particulier à mes parents et à mes sœurs, dont le soutien constant et les encouragements m'ont apporté la force et la motivation nécessaires pour mener à bien ce stage. Votre confiance en moi signifie plus que les mots ne peuvent exprimer.
Merci à toutes les personnes qui ont joué un rôle dans mon parcours au cours de ce stage.

Résumé

L'objectif principal de ce stage était de modéliser la valeur à vie du client (CLV) pour les clients professionnels et les entreprises associées à BNP Paribas sur les cinq prochaines années. Cette modélisation visait à prédire la valeur à long terme des clients après une Relation Économique Étendue (EER) avec BNP Paribas, en les segmentant en 10 classes distinctes allant de fort potentiel à nul.

Pour atteindre cet objectif, une gamme de techniques standard d'apprentissage automatique et de prétraitement a été employée. Un outil notable utilisé dans ce projet était le modèle AGBoost, une variation de XGBoost développée par BNP Paribas qui fournit une sortie linéaire, parmi d'autres modèles inspirés par divers articles académiques. L'application de ces méthodologies visait à améliorer la précision prédictive de la CLV et à fournir des insights exploitables pour la prise de décision stratégique de la banque.
Les résultats de cette étude devraient contribuer de manière significative à la compréhension des comportements des clients et à l'optimisation des stratégies d'engagement client chez BNP Paribas.

Abstract

The primary objective of this internship was to model the customer lifetime value (CLV) for professional clients and companies associated with BNP Paribas over the next five years. This modeling aimed to predict the long-term value of clients following an Extended Economic Relationship (EER) with BNP Paribas, segmenting them into 10 distinct classes ranging from high prospect to null.

To achieve this objective, a range of standard machine learning and preprocessing techniques were employed. A notable tool used in this project was the AGBoost model, a variation of XGBoost developed by BNP Paribas that provides a linear output, among other models inspired by various academic papers. The application of these methodologies aimed to enhance the predictive accuracy of CLV and provide actionable insights for the bank's strategic decision-making.

The outcomes of this study are expected to contribute significantly to understanding customer behaviors and optimizing client engagement strategies at BNP Paribas.

Chapter 1

1.1 Introduction

1.1.1 Aperçu de BNP Paribas

BNP Paribas est l'un des plus grands groupes bancaires au monde et un acteur majeur dans le secteur des services financiers, avec son siège social à Paris, France. Fondée en 1848, la banque a évolué pour devenir une institution financière internationale, offrant une gamme complète de services bancaires et financiers à une clientèle diversifiée, comprenant des particuliers, des entreprises, des institutions financières et des gouvernements [9].

Le groupe est présent dans 71 pays et emploie plus de 190 000 collaborateurs, dont plus de 145 000 en Europe. Avec une solide implantation en Europe, notamment en France, en Belgique, en Italie et au Luxembourg, BNP Paribas est également un acteur clé en Amérique du Nord, en Asie-Pacifique, au Moyen-Orient et en Afrique [9].

Les principaux domaines d'activité de BNP Paribas sont les suivants :

1. **Banque de détail et services** : BNP Paribas gère un vaste réseau de succursales à travers l'Europe, l'Asie, le Moyen-Orient et l'Afrique. Cette division offre une large gamme de produits et services financiers aux particuliers, notamment des comptes bancaires, des prêts, des assurances, des produits de placement, et des services de gestion de patrimoine [10]. En outre, la banque propose des solutions numériques innovantes pour améliorer l'expérience client et faciliter l'accès aux services financiers [8].

2. **Banque de financement et d'investissement (CIB)** : Cette division est dédiée aux entreprises multinationales, aux institutions financières et aux clients institutionnels. Elle offre une gamme complète de services financiers, y compris le conseil en fusion et acquisition, le financement structuré, la gestion d'actifs, et les services de titres. La CIB de BNP Paribas est reconnue pour son expertise en matière de financement durable, d'émission

d'obligations vertes, et de conseil en investissement responsable [7].

La banque met l'accent sur la transformation numérique et la durabilité, cherchant à intégrer des critères environnementaux, sociaux et de gouvernance (ESG) dans ses opérations et ses offres [11].

1.1.2 Aperçu de BCEF

La Banque de Crédit pour l'Économie Française (BCEF) est une division stratégique au sein de BNP Paribas, axée sur le soutien à l'économie française à travers des solutions de financement sur mesure. La BCEF joue un rôle essentiel dans le financement des petites et moyennes entreprises (PME) et des entreprises de taille intermédiaire (ETI), qui sont le moteur de l'économie française [9].

1.1.3 Introduction au Département EMC2 et à la Division EDA

Le département EMC2 de BNP Paribas, qui signifie "Ingénierie, Modélisation, Calculs et Conseil", est un pôle d'excellence en matière de science des données et d'analyse quantitative. Ce département joue un rôle crucial dans la transformation numérique de BNP Paribas, en fournissant des analyses avancées et des solutions basées sur les données pour soutenir les décisions stratégiques à travers le groupe [11].

La Division EDA (Exploratory Data Analysis) au sein d'EMC2 se spécialise dans l'exploration de grands ensembles de données pour découvrir des modèles cachés, des corrélations, et des insights précieux. L'analyse exploratoire des données est une étape cruciale dans le processus de modélisation, permettant de comprendre les structures sous-jacentes des données et d'informer le développement de modèles prédictifs robustes [8].

1.1.4 Objectif du Stage

L'objectif principal de ce stage était de développer un modèle prédictif pour estimer la valeur à vie des clients (Customer Lifetime Value - CLV) pour les clients professionnels et les entreprises associées à BNP Paribas sur une période de cinq ans. La CLV est une métrique clé qui permet à la banque de segmenter sa clientèle en fonction de leur valeur potentielle, ce qui est essentiel pour la mise en œuvre de stratégies marketing ciblées et l'optimisation des ressources allouées aux différentes catégories de clients [8].

Défis rencontrés : Le développement de ce modèle a impliqué plusieurs défis, notamment le traitement de données asymétriques et inflationnées

(où une grande proportion des valeurs de la variable cible sont nulles), la sélection des caractéristiques pertinentes, et l'optimisation des algorithmes de modélisation pour améliorer la précision des prédictions [11].

En résumé, ce stage a permis de développer un cadre analytique robuste pour l'estimation de la CLV, offrant à BNP Paribas des outils précieux pour la gestion de la relation client et la prise de décision stratégique à long terme [9].

1.2 Revue de la Littérature

1.2.1 Introduction à la Valeur à Vie du Client (CLV)

La Valeur à Vie du Client (Customer Lifetime Value - CLV) est une métrique clé dans la gestion de la relation client (CRM), utilisée pour estimer le revenu net qu'une entreprise peut attendre d'un client tout au long de sa relation commerciale [3]. La CLV permet aux entreprises de segmenter leur clientèle en fonction de leur valeur potentielle, d'optimiser les dépenses marketing, et de prendre des décisions éclairées sur les stratégies de fidélisation [14].

Dans le secteur bancaire, la CLV est particulièrement pertinente en raison de la nature récurrente des transactions financières et de l'importance de la fidélité des clients pour la rentabilité à long terme. Les banques utilisent la CLV pour identifier les clients à forte valeur ajoutée, ajuster leurs offres de produits et améliorer l'expérience client [15]. De plus, la CLV aide à identifier les risques de perte de clients et à développer des stratégies de rétention pour les segments de clients à haut risque [4].

La CLV est traditionnellement calculée à l'aide de modèles statistiques tels que le modèle RFM (Récence, Fréquence, Montant), qui se base sur les comportements transactionnels passés des clients. Cependant, avec l'avènement de l'apprentissage automatique et des techniques de traitement de grandes quantités de données, des approches plus sophistiquées et plus précises ont été développées pour modéliser la CLV

1.2.2 Approches d'Apprentissage Automatique pour la Modélisation de la CLV

Les approches d'apprentissage automatique pour la modélisation de la CLV sont devenues de plus en plus populaires en raison de leur capacité à traiter de grandes quantités de données et à identifier des modèles complexes dans les comportements des clients [6]. Contrairement aux méthodes traditionnelles, les techniques d'apprentissage automatique peuvent intégrer une multitude

de variables, telles que les interactions numériques des clients, les données démographiques, et les comportements d'achat, pour améliorer la précision des prédictions de la CLV [12].

1. Modèles de Régression : Les modèles de régression, tels que la régression linéaire, la régression logistique, et la régression de Poisson, sont couramment utilisés pour estimer la CLV en fonction de variables explicatives multiples [13]. Ces modèles permettent de quantifier l'impact de différentes variables sur la valeur à vie d'un client, facilitant ainsi la segmentation des clients et l'optimisation des stratégies marketing.

2. Modèles d'Arbres de Décision : Les arbres de décision, tels que les arbres de régression et les forêts aléatoires, sont des techniques populaires pour la modélisation de la CLV en raison de leur capacité à capturer les interactions non linéaires entre les variables explicatives [1]. Les forêts aléatoires, en particulier, ont montré une grande efficacité dans la prédiction de la CLV, car elles réduisent le risque de surapprentissage et améliorent la généralisation du modèle [5].

3. Modèles d'Ensemble : Les modèles d'ensemble, tels que le Gradient Boosting et l'AGBoost, une variation développée par BNP Paribas, combinent plusieurs modèles de base pour améliorer la précision prédictive et réduire l'erreur de généralisation [2]. Ces modèles sont particulièrement adaptés à la modélisation de la CLV, car ils peuvent gérer des distributions de données complexes et des variables multicolinéaires.

4. Réseaux de Neurones et Apprentissage Profond : Les réseaux de neurones, et plus récemment, les réseaux de neurones profonds (Deep Learning), ont été utilisés pour modéliser la CLV dans des contextes où les données sont hautement dimensionnelles et non structurées [16]. Ces modèles peuvent capturer des relations complexes entre les variables et sont particulièrement efficaces pour les grandes bases de données transactionnelles.

5. Modèles Bayésiens : Les approches bayésiennes, telles que le modèle de mélange bayésien, permettent de modéliser l'incertitude et de prendre en compte les distributions a priori dans la prédiction de la CLV [13]. Ces modèles sont utiles dans des situations où les données sont rares ou où il existe une forte hétérogénéité entre les clients.

En conclusion, l'utilisation de l'apprentissage automatique pour la modélisation de la CLV représente une avancée significative dans la capacité des entreprises à comprendre et à anticiper les comportements des clients. Les modèles ML offrent des prédictions plus précises et des insights plus approfondis, permettant ainsi aux entreprises de maximiser la valeur à long terme de leurs clients [6].

Chapter 2

2.1 Traitement et Analyse des Données

2.1.1 Introduction aux Données

Le jeu de données utilisé dans cette étude provient de plusieurs DataFrames, appelés profils de stock, chacun représentant une capture d'image des données enregistrées pour un mois donné. En raison de la nature confidentielle des données, une description détaillée de certaines variables ne peut être fournie. Cependant, le jeu de données contient à la fois des variables catégorielles et continues, avec une attention particulière portée à la prédiction de la **Valeur à Vie du Client (CLV)** sur plusieurs années.

Le jeu de données couvre la période de 2018 à 2023, et a été agrégé afin de calculer la **Valeur Cumulative de la CLV (CCLV)**, qui correspond à la somme des valeurs annuelles de la CLV sur les six ans considérés. Cette variable CCLV représente la variable cible de notre analyse et de notre modélisation. La Figure 2.1 illustre visuellement le processus d'agrégation des différents profils de stock.

Aperçu des Variables

Le jeu de données comprend une variété de variables catégorielles et continues. Les variables continues incluent des indicateurs financiers clés tels que le flux annuel et des caractéristiques de performance des clients sur plusieurs périodes. Quant aux variables catégorielles, elles concernent des attributs descriptifs des clients tels que leur secteur d'activité et leur localisation géographique. La variable cible principale est la **CCLV**, qui est une somme des valeurs CLV annuelles agrégées sur plusieurs années.

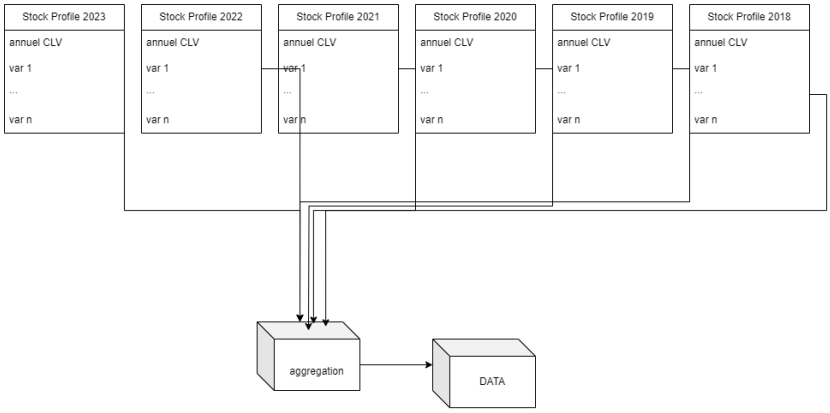

Figure 2.1: Diagramme du processus d'agrégation des Profils de Stock

Observations Initiales et Défis

Lors de l'analyse initiale des variables continues, nous avons observé une distribution fortement asymétrique à droite (distribution étirée vers les valeurs élevées) pour la plupart des variables. Cette asymétrie présente un défi pour les modèles statistiques et d'apprentissage automatique, car elle peut fausser les résultats et les conclusions. De plus, un grand nombre de valeurs aberrantes ont été identifiées, notamment dans les variables en lien avec la CLV, ce qui pourrait affecter les performances du modèle si elles ne sont pas correctement traitées.

Les variables catégorielles, bien que moins sujettes aux effets des valeurs aberrantes, présentent un déséquilibre marqué dans certaines catégories, avec certaines classes surreprésentées par rapport à d'autres.

En plus des valeurs aberrantes, des données manquantes ont été détectées dans certaines variables, nécessitant des méthodes d'imputation ou des solutions adaptées pour assurer l'intégrité de l'analyse ultérieure. Ces défis initiaux de gestion des valeurs manquantes, de valeurs aberrantes et de distributions biaisées ont été les premières étapes du prétraitement des données.

Défis spécifiques liés aux Données Manquantes et Asymétrie

- **Valeurs Manquantes** : Plusieurs variables catégorielles présentaient des données manquantes. Pour ces variables, les valeurs manquantes ont été imputées en les classant dans une catégorie spéciale "manquante". En revanche, pour les variables continues, aucune valeur manquante n'a été détectée, ce qui a simplifié le traitement de ces variables continues.

- **Asymétrie des Données** : Les variables continues présentaient des distributions extrêmement asymétriques, rendant nécessaire l'application de transformations pour réduire cette asymétrie et améliorer la normalité des distributions. Des méthodes telles que la winsorisation ou la transformation logarithmique ont été envisagées pour stabiliser les distributions et minimiser l'effet des valeurs extrêmes.

2.2 Nettoyage des Données

Le processus de nettoyage des données a été une étape cruciale pour garantir la qualité des données avant leur utilisation dans la modélisation. Dans cette section, nous détaillons les méthodes utilisées pour traiter les valeurs manquantes et les valeurs aberrantes, ainsi que la validation des données après nettoyage.

2.2.1 Gestion des Valeurs Manquantes

Dans notre jeu de données, nous avons observé des valeurs manquantes principalement dans les variables catégorielles. Pour ces variables, les valeurs manquantes ont été imputées en créant une nouvelle catégorie missing , afin de préserver l'intégrité de l'ensemble des données. Concernant les variables continues, aucune valeur manquante n'a été observée, ce qui a facilité l'analyse de ces variables sans nécessiter d'imputation supplémentaire.

2.2.2 Traitement des Valeurs Aberrantes

Les valeurs aberrantes représentent un défi majeur, notamment en raison de la forte asymétrie observée dans les variables continues. Pour traiter ces valeurs aberrantes, plusieurs niveaux de winsorisation ont été appliqués, limitant ainsi l'effet des extrêmes tout en conservant la structure sous-jacente des données.

Les figures ci-dessus illustrent l'évolution de la distribution des données après l'application de la winsorisation aux niveaux de 0,01, 0,05 et 0,1. Il est clairement visible que les niveaux de winsorisation plus élevés réduisent progressivement l'effet des valeurs aberrantes tout en maintenant la majorité des observations dans leur plage d'origine.

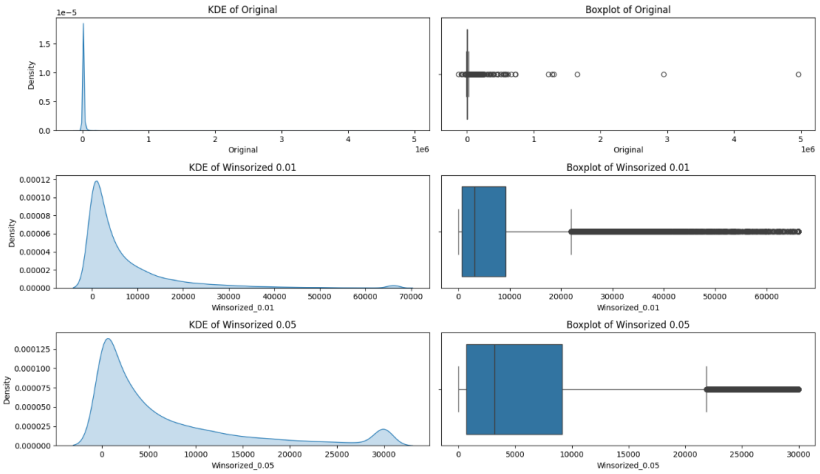

Figure 2.2: Distribution de la variable avant et après winsorisation à différents niveaux.

2.2.3 Approche Basée sur le Clustering pour la Détection des Valeurs Aberrantes

En complément de la winsorisation, une approche de clustering a été explorée pour identifier les valeurs aberrantes de manière plus robuste. Cette méthode nous a permis de segmenter les données en deux groupes principaux : les valeurs normales et les valeurs considérées comme aberrantes. Une fois ces valeurs aberrantes identifiées, elles ont été remplacées par le 95ème percentile du groupe des valeurs normales.

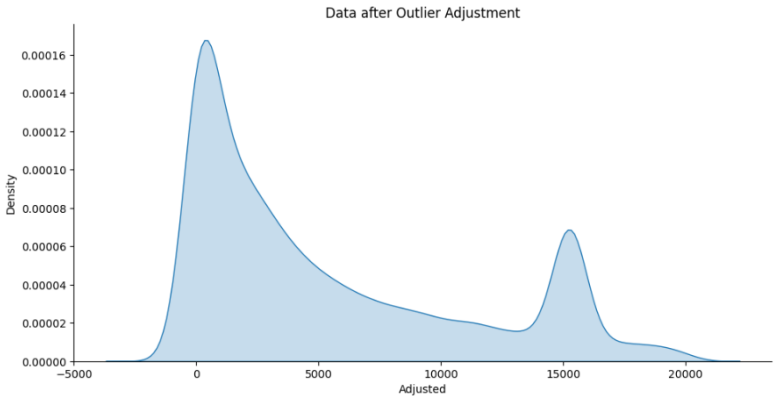

Figure 2.3: Données après ajustement des valeurs aberrantes.

La figure ci-dessus montre la distribution après l'ajustement des valeurs aberrantes à l'aide de la méthode de clustering. La distribution résultante

présente moins de variabilité extrême, permettant ainsi une modélisation plus stable.

2.2.4 Validation des Données Après Nettoyage

Une fois le processus de nettoyage terminé, une validation des données a été effectuée pour s'assurer de la cohérence des transformations. Nous avons utilisé des méthodes de visualisation, telles que les courbes de densité et les boxplots, pour évaluer l'impact des transformations appliquées.

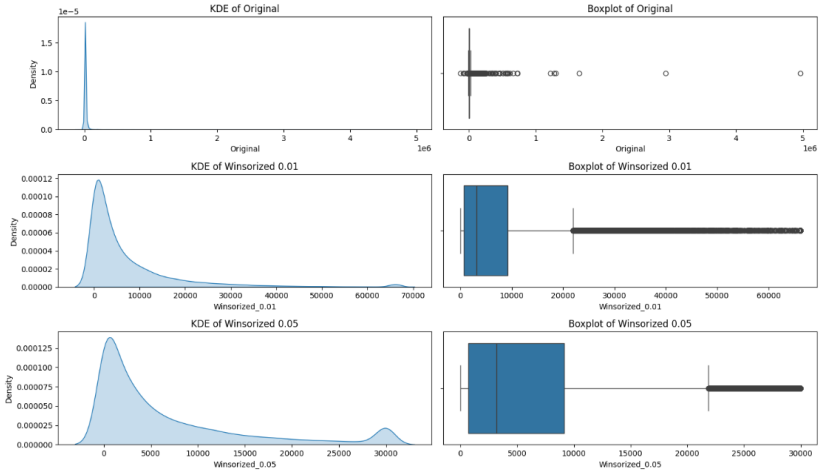

Figure 2.4: Données originales avec surbrillance des valeurs aberrantes avant et après ajustement.

Les graphiques montrent la distribution des données originales avec des valeurs aberrantes mises en évidence, ainsi que la distribution après ajustement. Cette approche nous a permis de conserver l'intégrité des données tout en atténuant l'effet des valeurs extrêmes, ce qui améliore la fiabilité des résultats lors des étapes de modélisation.

En conclusion, le processus de nettoyage des données, incluant la gestion des valeurs manquantes et des valeurs aberrantes par winsorisation et clustering, a significativement amélioré la qualité des données. Ces ajustements ont permis d'obtenir des données prêtes pour une modélisation robuste, garantissant une meilleure performance des modèles prédictifs dans les étapes suivantes.

2.3 Transformation des Données

Le processus de transformation des données est une étape essentielle pour préparer les données avant l'application des modèles de machine learning. Cette section présente les techniques utilisées pour encoder les variables catégorielles, normaliser les variables numériques et corriger l'asymétrie des variables continues. Enfin, nous évaluons la normalité des distributions à l'aide de tests statistiques.

2.3.1 Encodage des Variables Catégorielles

Les variables catégorielles, qui représentent des attributs qualitatifs, doivent être encodées sous forme numérique pour pouvoir être utilisées dans les algorithmes de machine learning. Deux méthodes principales d'encodage ont été utilisées dans cette étude :

- **Encodage One-Hot**: Pour les variables ayant un faible nombre de catégories uniques, l'encodage one-hot a été appliqué. Cette méthode crée une nouvelle colonne pour chaque modalité de la variable catégorielle, attribuant la valeur 1 ou 0 selon que la modalité est présente ou non dans chaque observation.
- **Encodage Ordinal**: Pour les variables catégorielles ordinales, c'est-à-dire les variables dont les modalités ont un ordre naturel, un encodage ordinal a été utilisé, attribuant un entier à chaque modalité en fonction de son rang.

Ces techniques ont permis de conserver les informations des variables catégorielles tout en les rendant compatibles avec les algorithmes de modélisation.

2.3.2 Mise à l'Échelle et Normalisation des Variables Numériques

Les variables numériques, en particulier celles avec des échelles de valeur très différentes, peuvent poser des problèmes lors de la modélisation. Afin d'éviter que certaines variables dominent les autres, des techniques de mise à l'échelle et de normalisation ont été utilisées :

- **Mise à l'échelle min-max**: Cette méthode a été utilisée pour ramener toutes les variables numériques dans une plage entre 0 et 1, facilitant la convergence des algorithmes d'apprentissage.

- **Normalisation Z-score**: Les variables ont également été transformées en scores Z, en soustrayant la moyenne et en divisant par l'écart-type. Cela permet de centrer les données autour de 0 avec un écart-type de 1.

Ces techniques assurent une distribution plus uniforme des variables numériques, réduisant les biais induits par des échelles différentes.

2.3.3 Gestion de l'Asymétrie des Variables Continues

De nombreuses variables continues dans notre jeu de données présentaient une distribution fortement asymétrique à droite, comme illustré dans les figures ci-dessous. Pour rendre ces distributions plus symétriques, plusieurs transformations ont été testées :

- **Transformation Logarithmique**: Appliquée aux variables avec des valeurs strictement positives, cette transformation réduit l'impact des grandes valeurs en compressant la queue droite de la distribution.
- **Transformation de Box-Cox**: Utilisée pour les variables positives, cette transformation permet d'améliorer la normalité des distributions en ajustant les données selon un paramètre lambda.
- **Transformation de Yeo-Johnson**: Contrairement à la transformation logarithmique, la méthode Yeo-Johnson peut être appliquée aux variables avec des valeurs positives et négatives, offrant une plus grande flexibilité.

La figure montre l'effet de chaque transformation sur une variable continue asymétrique, avec une amélioration visible de la symétrie des distributions.

2.3.4 Tests de Normalité

Une fois les transformations appliquées, des tests de normalité ont été effectués pour évaluer dans quelle mesure les distributions résultantes se rapprochaient d'une distribution normale. Les tests suivants ont été utilisés :

- **Test de Shapiro-Wilk**: Ce test statistique vérifie si une donnée suit une distribution normale. Une p-valeur inférieure à 0,05 indique que la distribution est significativement différente de la normalité.

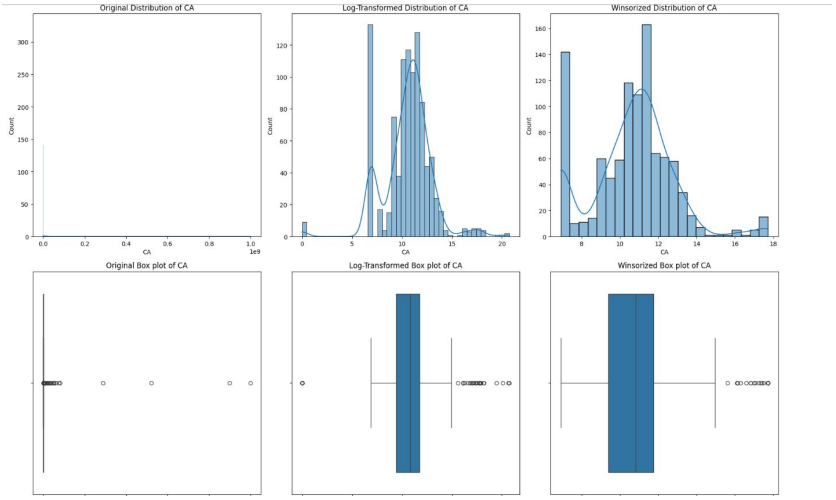

Figure 2.5: Comparaison des différentes transformations appliquées à une variable continue.

- **Test de Kolmogorov-Smirnov**: Ce test compare la distribution observée à une distribution normale théorique. Comme pour le test de Shapiro-Wilk, une p-valeur inférieure à 0,05 rejette l'hypothèse de normalité.

Bien que certaines variables ne suivent toujours pas une distribution normale parfaite, ces transformations ont réduit l'asymétrie des distributions, rendant les données mieux adaptées à la modélisation.

2.3.5 Analyse Bivariée (Corrélations entre les Variables Explicatives et la Variable Cible)

L'analyse bivariée vise à explorer les relations entre les variables explicatives et la variable cible, ici **total_pnb**, qui représente la **Valeur Cumulative de la CLV (CCLV)**. Nous avons évalué les corrélations linéaires et non linéaires entre les variables continues et la variable cible pour comprendre les relations sous-jacentes et identifier les potentiels problèmes de multicolinéarité.

Matrice de Corrélation

Nous avons utilisé une matrice de corrélation pour examiner les relations entre les variables explicatives et **total_pnb**. Cette matrice fournit des coefficients de corrélation de Pearson, qui mesurent l'intensité et la direction

des relations linéaires entre deux variables. Les corrélations avec des valeurs proches de -1 ou 1 indiquent une forte relation, tandis que les valeurs proches de 0 indiquent peu ou pas de relation.

La figure 2.6 montre la matrice de corrélation pour les principales variables continues.

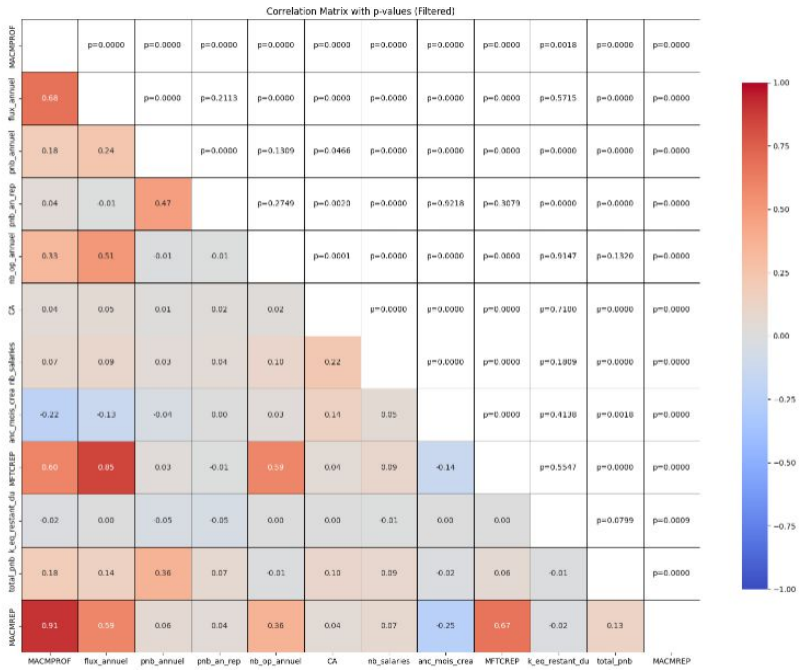

Figure 2.6: Matrice de corrélation montrant les relations entre les variables explicatives et **total_pnb**.

Observations principales :

- Les variables continues telles que **flux_annuel** et **pnb_annuel** présentent des corrélations modérées à fortes avec **total_pnb** (coefficients respectifs de 0,60 et 0,50). Cela indique une relation positive et significative entre ces variables et la variable cible. Cependant, ces deux variables observent la performance annuelle du client, ce qui diffère de notre objectif centré sur les clients deux mois après la **EER**.

- La variable **MACNPROF** présente une corrélation très forte avec **flux_annuel** (0,91), ce qui indique une multicolinéarité entre ces deux variables. Une telle redondance peut nuire à la qualité du modèle en introduisant des informations similaires plusieurs fois.

- Plusieurs autres variables ont montré des relations plus faibles avec **total_pnb**, telles que **nb_salaries** (corrélation de 0,12), et ont

donc été conservées dans le modèle pour fournir une diversité d'informations sans créer de multicolinéarité excessive.

Décisions basées sur des Considérations Statistiques et Métiers

En plus de l'analyse des corrélations, des décisions ont été prises en tenant compte à la fois des résultats statistiques et des considérations métiers. Il est important de noter que certaines variables fortement corrélées avec **total_pnb** ont été exclues du modèle final pour des raisons liées à la nature du projet, plutôt que pour des raisons purement statistiques.

- **Exclusion de flux_annuel et pnb_annuel** : Bien que ces variables montrent une corrélation significative avec **total_pnb**, elles sont des indicateurs annuels, et donc ne correspondent pas directement à notre objectif d'étudier la performance des clients deux mois après l'EER. Étant donné que le projet se concentre sur cette période spécifique, ces variables ont été retirées du modèle final pour éviter de capter des informations non pertinentes.

- **Gestion de la Multicolinéarité** : La corrélation très forte entre **MACNPROF** et **flux_annuel** a mis en évidence une redondance potentielle dans les informations capturées par ces deux variables. Sur la base de discussions métiers et des résultats de l'analyse, nous avons décidé de conserver **MACNPROF** dans le modèle et de retirer **flux_annuel**. Bien que cette décision ait été en partie guidée par la corrélation statistique, elle reflète également une compréhension approfondie des besoins métiers.

Corrélation entre Variables Discrètes

Pour les variables numériques discrètes, des méthodes spécifiques ont été appliquées pour explorer les relations avec **total_pnb**. Nous avons utilisé les métriques suivantes :

- **Corrélation de Spearman** : Cette métrique permet de mesurer la force et la direction de la relation monotone entre deux variables, sans supposer de relation linéaire. Elle est particulièrement utile lorsque les relations entre les variables ne suivent pas une tendance strictement linéaire.

- **Tau de Kendall** : Mesure la force de l'association entre deux variables ordinales, et fournit une approche alternative à la corrélation de Pearson, adaptée aux variables discrètes.

- **Information mutuelle** : Évalue la quantité d'information partagée entre deux variables. Elle permet d'identifier les relations non linéaires potentielles entre les variables discrètes et la cible.

L'analyse a révélé des relations intéressantes entre plusieurs variables discrètes et **total_pnb**.

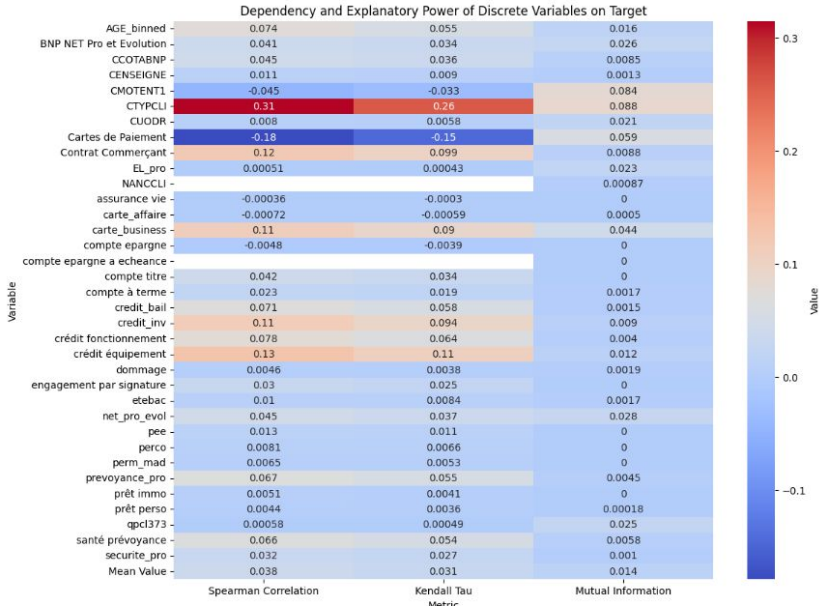

Figure 2.7: Carte de chaleur des corrélations de Kendall Tau pour les variables discrètes et **total_pnb**.

Observations sur les Variables Discrètes :

- **CTYPCLI** présente une corrélation notable avec **total_pnb** (tau = 0,31), indiquant que cette variable a un pouvoir explicatif non négligeable concernant la cible. Ce type de client a donc été conservé pour les étapes de modélisation ultérieures.

- **CMOTENT1** montre des relations modérées avec la cible (tau = 0,26), justifiant également son inclusion dans le modèle.

- D'autres variables discrètes ont montré des corrélations plus faibles, comme **Carte de Paiement**, qui présente une corrélation négative avec **total_pnb** (tau = -0,15), suggérant que ces variables peuvent ne pas être des prédicteurs puissants.

Corrélation entre Variables Catégorielles

Les variables catégorielles, étant de nature qualitative, nécessitent des méthodes différentes pour évaluer leur relation avec **total_pnb**. Nous avons utilisé les métriques suivantes :

- **V de Cramér** : Cette métrique mesure la force de l'association entre deux variables catégorielles. Elle est particulièrement utile pour analyser les relations entre les catégories et la variable cible continue.
- **Eta carré (^{2})** : Cet indicateur permet d'évaluer la proportion de la variance de **total_pnb** qui est expliquée par une variable catégorielle.

Les résultats de cette analyse ont montré que certaines variables catégorielles avaient un pouvoir explicatif limité, bien qu'elles contribuent toujours à la compréhension du modèle global.

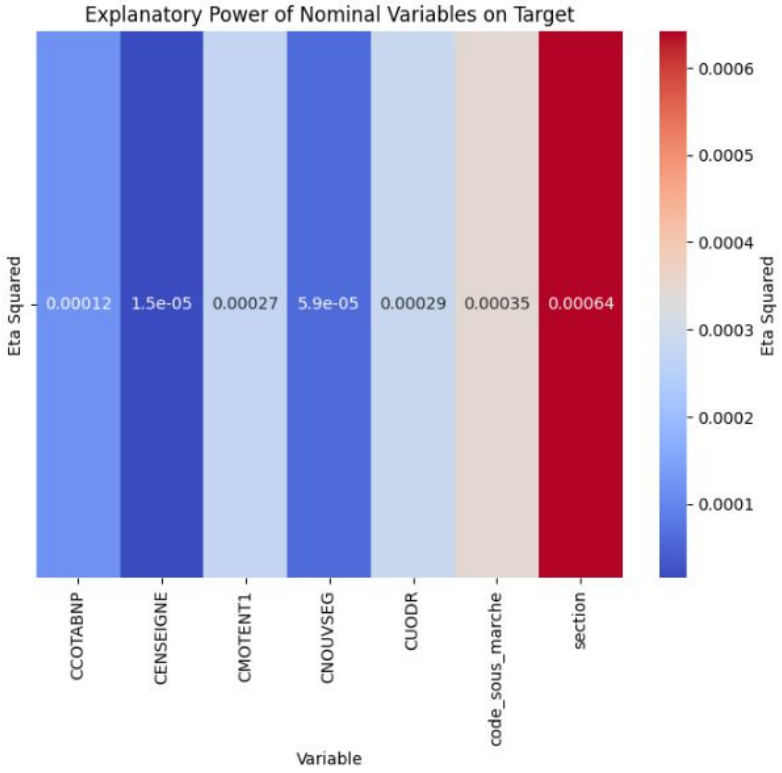

Figure 2.8: Pouvoir explicatif des variables catégorielles sur **total_pnb** mesuré par l'eta carré.

Observations sur les Variables Catégorielles :

- La variable **section** a montré un pouvoir explicatif modéré avec une valeur d'eta carré de 0,00064, suggérant qu'elle capture une petite part de la variance dans **total_pnb**. Cela justifie son inclusion dans les modèles prédictifs.
- **CTYPCLI** et **CUODR** ont également montré des relations intéressantes avec **total_pnb** en termes de V de Cramér, renforçant leur importance pour le modèle.
- Les autres variables catégorielles, comme **code_sous_marche**, ont montré un pouvoir explicatif plus faible et peuvent ne pas contribuer de manière significative à la modélisation de la CLV après l'EER.

Conclusion de l'Analyse Bivariée

L'analyse bivariée a permis de mettre en évidence les relations clés entre les variables explicatives (continues, discrètes, et catégorielles) et la variable cible **total_pnb**. Les décisions de conservation ou d'exclusion des variables ont été prises sur la base de considérations statistiques et métiers. Les variables comme **CTYPCLI**, **MACNPROF**, et **section** se sont révélées significatives, tant sur le plan statistique que métier, et seront donc incluses dans les modèles prédictifs finaux pour l'estimation de la CLV après l'EER.

2.3.6 Conclusion

Les étapes de transformation des données, incluant l'encodage des variables catégorielles, la normalisation des variables numériques et la correction de l'asymétrie des variables continues, ont permis de préparer un jeu de données propre et équilibré pour la modélisation. Les tests de normalité ont confirmé que les transformations appliquées avaient amélioré la distribution des variables continues, renforçant ainsi la robustesse des modèles prédictifs.

Chapter 3

3.1 Modélisation et Implémentation

3.1.1 Développement du Modèle Initial

Dans cette section, nous décrivons les modèles de machine learning initiaux utilisés, les étapes de prétraitement des données, les techniques d'ingénierie des features, ainsi que les premières métriques de performance. Nous abordons également les défis rencontrés au cours de cette phase initiale.

Choix du Modèle

Le choix des modèles **XGBoost** et **AGBoost** a été principalement dicté par les contraintes de production de notre environnement. En effet, **AGBoost** est un modèle développé par le PNB, construit sur la base de **XGBoost**. C'est un modèle d'ensemble qui produit une sortie linéaire, ce qui impose une certaine convergence de toutes les approches et modèles vers **AGBoost**.

XGBoost a été utilisé comme solution temporaire pour l'exploration et l'optimisation des hyperparamètres, étant donné qu'il partage l'espace des hyperparamètres avec **AGBoost**. Ainsi, **XGBoost** nous a permis de réaliser des tests rapides, des ajustements et des validations préliminaires avant l'intégration complète dans **AGBoost** pour la phase de production.

Prétraitement des Données et Ingénierie des Features

Avant l'entraînement des modèles, plusieurs étapes de prétraitement des données ont été effectuées pour améliorer la qualité des données et maximiser la performance des modèles.

Transformation Logarithmique : Nous avons appliqué une transformation logarithmique à certaines variables présentant une distribution fortement asymétrique à droite, afin de stabiliser leur distribution et réduire l'impact des valeurs extrêmes.

Gestion des Outliers : Les outliers identifiés lors de l'analyse exploratoire ont été supprimés du DataFrame avant l'entraînement des modèles. Cela a permis de réduire l'influence des points de données aberrants sur les prédictions.

Valeurs Manquantes : La gestion des valeurs manquantes avait déjà été abordée dans les phases de prétraitement précédentes, et aucune autre imputation n'a été nécessaire avant l'entraînement.

Ingénierie des Features : Nous avons tenté de créer des features supplémentaires via des transformations polynomiales sur certaines variables, mais ces transformations n'ont pas amélioré les performances des modèles. En conséquence, elles ont été abandonnées pour les itérations suivantes.

Métriques de Performance Initiales

Étant donné que notre problème est de nature régressive, les principales métriques utilisées pour évaluer les modèles ont été le **R^2** et la **RMSE** (Root Mean Squared Error). Ces deux métriques ont permis d'évaluer la qualité des prédictions et de comparer les performances des différents modèles.

Problèmes rencontrés : Dans les premières itérations, nous avons constaté quelques cas de surapprentissage (**overfitting**), principalement en raison de la petite taille de notre jeu de données. Ce problème a été corrigé en ajustant certains hyperparamètres, en appliquant des régularisations appropriées et en augmentant la pénalisation dans le modèle.

Contraintes Computationnelles : Aucune contrainte computationnelle majeure n'a été rencontrée, en raison de la taille relativement modeste de notre jeu de données. Les temps de calcul pour l'entraînement des modèles sont restés raisonnables tout au long du processus, permettant une optimisation itérative efficace.

3.2 Recherche pour l'Amélioration des Modèles

Dans le cadre de ce projet, plusieurs approches théoriques et expérimentales ont été mises en œuvre pour surmonter les défis de la prédiction de la valeur à vie du client (CLV). En particulier, les problématiques liées à l'asymétrie des données et à la prédiction de la valeur à vie pour les clients B2B ont exigé une exploration approfondie des modèles de régression adaptés aux distributions longues queues et aux données fortement asymétriques.

3.2.1 Quadratic Mean Based Supervised Learning Model for Managing Data Skewness

The Quadratic Mean Learning (QMLearn) framework was implemented to address the issue of skewed data distributions, which are common in many real-world datasets. The QMLearn method adjusts the empirical risk minimization by employing the quadratic mean rather than the arithmetic mean, allowing the model to handle data skewness more robustly.

Introduction du Cadre QMLearn

L'approche QMLearn modifie la définition traditionnelle du risque empirique en utilisant la moyenne quadratique pour mieux gérer les distributions asymétriques. La fonction de risque empirique pour QMLearn est définie comme suit :

$$R_{emp}^{Q}(w) = \sqrt{\frac{\left(\frac{\sum_{i=1}^{n_1} l(x_i,y_i,w)}{n_1}\right)^2 + \left(\frac{\sum_{i=n_1+1}^{n} l(x_i,y_i,w)}{n_2}\right)^2}{2}},$$

où n_1 et n_2 sont les nombres d'exemples pour chaque classe, et $l(x_i, y_i, w)$ représente la fonction de perte (telle que l'erreur quadratique moyenne) entre l'exemple x_i et le label réel y_i. Cette méthode égalise l'influence des deux classes sur l'apprentissage du modèle, ce qui permet de mieux gérer les distributions biaisées.

Fonction de Perte Quadratique Simplifiée

Pour simplifier l'implémentation, la fonction de perte quadratique peut être reformulée à l'aide de deux termes A et B représentant la somme des erreurs dans les deux groupes de données :

$$A = \frac{2}{n}\sum_{i=1}^{\frac{n}{2}}(\hat{y}_i - y_i)^2, \quad B = \frac{2}{n}\sum_{i=\frac{n}{2}+1}^{n}(\hat{y}_i - y_i)^2.$$

Avec ces définitions, la fonction de perte devient :

$$R^Q_{emp}(w) = \sqrt{\frac{A^2 + B^2}{2}}.$$

Calcul des Gradients et Hessiennes

L'optimisation dans XGBoost nécessite le calcul des gradients et des hessiennes de la fonction de perte quadratique. Ces dérivées permettent d'ajuster les paramètres du modèle lors de l'entraînement.

Le gradient de la fonction de perte quadratique est défini par :

$$\text{grad}_i = \frac{1}{2} \cdot (R^Q_{emp}(w))^{-0.5} \cdot C_i \cdot C'_i,$$

où C_i est égal à A si $i \leq \frac{n}{2}$, et B sinon, et $C'_i = \frac{4(\hat{y}_i - y_i)}{n}$.

La hessienne est donnée par :

$$\text{hess}_i = \left(-\frac{1}{4} \cdot (R^Q_{emp}(w))^{-1.5} \cdot C_i \cdot C'_i \right) + \frac{1}{2} \cdot (R^Q_{emp}(w))^{-0.5} \cdot \left((C'_i)^2 + \frac{4C_i}{n} \right).$$

Ces formules permettent à XGBoost d'optimiser le modèle en utilisant la moyenne quadratique pour mieux traiter les ensembles de données fortement biaisés.

Stratégie d'Implémentation

L'implémentation de cette fonction de perte personnalisée a été intégrée dans XGBoost en tant qu'objectif pour gérer les tâches de régression sur des données asymétriques. En définissant correctement la fonction de perte quadratique, nous avons pu ajuster le modèle pour tenir compte des distributions biaisées et améliorer les prédictions pour des variables fortement asymétriques.

3.3 Modèle Probabiliste Profond pour la Prédiction de la CLV

Cette section résume les concepts clés et les solutions proposées dans l'article "A Deep Probabilistic Model for Customer Lifetime Value Prediction" par Xiaojing Wang, Tianqi Liu, et Jingang Miao. L'article traite des défis liés à la prédiction de la **Valeur Vie Client (CLV)** dans des contextes de données très biaisées et gonflées de zéros. Les auteurs proposent une approche probabiliste innovante pour améliorer la précision et la robustesse des prédictions de la CLV.

3.3.1 Défis dans la Prédiction de la CLV

La prédiction de la CLV est essentielle pour les entreprises souhaitant estimer les revenus futurs potentiels de leurs clients. Cependant, elle présente des défis majeurs liés à la nature des données :

- **Distribution asymétrique**: Les données de la CLV sont souvent fortement biaisées à droite, où la majorité des clients génèrent peu ou pas de revenus, tandis qu'une minorité de clients à forte valeur génère des revenus disproportionnés.
- **Gonflement de zéros**: Une proportion importante des clients peut avoir une CLV égale à zéro, représentant des acheteurs ponctuels qui ne reviennent pas. Cela crée une inflation de zéros dans le jeu de données, compliquant le processus de modélisation.

Les modèles de régression classiques, tels que ceux utilisant la perte des moindres carrés (MSE), sont peu adaptés à ces défis, car ils sont sensibles aux valeurs extrêmes et supposent une distribution normale des erreurs.

3.3.2 Distribution ZILN (Zero-Inflated Lognormal)

Pour traiter la nature asymétrique et gonflée de zéros des données de la CLV, les auteurs proposent de modéliser la distribution de la CLV à l'aide d'une **distribution ZILN (Zero-Inflated Lognormal)**. Cette approche capture à la fois la probabilité qu'un client ait une CLV nulle et la variabilité parmi les clients ayant une CLV non nulle.

La distribution ZILN combine une *masse ponctuelle à zéro* (pour traiter l'inflation de zéros) avec une *distribution lognormale* (pour gérer la nature asymétrique des CLV non nulles). Ce modèle est particulièrement adapté à la prédiction de la CLV, car il offre un cadre flexible qui peut capturer les caractéristiques distinctes des données.

3.3.3 Fonction de Perte de la Distribution ZILN

Les auteurs dérivent une fonction de perte basée sur la vraisemblance négative d'une variable aléatoire distribuée selon une ZILN. Cette fonction de perte modélise efficacement les deux composantes principales de la CLV :

- **La probabilité de CLV nulle** (pour les clients qui ne feront pas d'autres achats).

- **La distribution lognormale des CLV non nulles** (pour les clients qui effectueront des achats supplémentaires).

La fonction de perte ZILN est définie comme suit :

$$L_{\text{ZILN}}(x; p, \mu, \sigma) = -\mathbf{1}_{\{x=0\}} \log(1-p) - \mathbf{1}_{\{x>0\}} \left(\log p - L_{\text{Lognormal}}(x; \mu, \sigma)\right)$$

où :

- x représente la CLV observée.
- p est la probabilité que la CLV soit non nulle.
- μ et σ sont la moyenne et l'écart-type de la distribution lognormale pour les CLV non nulles.
- $L_{\text{Lognormal}}(x; \mu, \sigma)$ est la vraisemblance négative de la distribution lognormale.

Cette fonction de perte permet au modèle d'apprendre à partir des CLV nulles et non nulles, capturant ainsi avec précision la distribution des données de la CLV.

3.3.4 Gestion de l'Asymétrie et de l'Inflation des Zéros

La fonction de perte ZILN offre plusieurs avantages clés pour traiter la nature asymétrique et gonflée de zéros des données de la CLV :

- **Gestion de l'inflation de zéros**: En modélisant directement la probabilité de CLV nulle, la fonction de perte ZILN gère efficacement la proportion élevée de zéros dans le jeu de données, offrant une représentation plus précise des acheteurs ponctuels.
- **Modélisation des queues lourdes**: La composante lognormale de la distribution ZILN capture la nature asymétrique et à queue lourde des CLV non nulles, permettant au modèle de tenir compte de la variabilité parmi les clients récurrents.
- **Flexibilité et robustesse**: La fonction de perte ZILN offre un cadre flexible pouvant être appliqué aux modèles linéaires et non linéaires, améliorant ainsi la robustesse et la performance en généralisation en présence de données biaisées.

3.3.5 Solution et Avantages de l'Approche ZILN

Les auteurs proposent d'utiliser la fonction de perte ZILN dans un cadre d'apprentissage profond pour tirer parti de sa flexibilité et de sa puissance de modélisation. Cette approche offre plusieurs avantages :

- **Modélisation unifiée**: La fonction de perte ZILN permet au modèle d'effectuer à la fois une classification (prédiction de la récurrence d'un client) et une régression (prédiction de la CLV des clients récurrents) dans un cadre unifié.
- **Amélioration de la précision des prédictions**: En modélisant avec précision la nature gonflée de zéros et asymétrique des données de la CLV, la fonction de perte ZILN améliore la précision des prédictions, notamment pour les clients à haute valeur.
- **Évolutivité et adaptabilité**: La nature probabiliste du modèle ZILN le rend évolutif pour de grands ensembles de données et adaptable à divers domaines présentant des caractéristiques de données similaires.

3.3.6 Conclusion

L'article introduit une approche probabiliste profonde pour la prédiction de la CLV à l'aide d'une nouvelle fonction de perte basée sur la distribution ZILN. En modélisant efficacement la nature gonflée de zéros et à queue lourde des données de la CLV, l'approche ZILN fournit une solution robuste permettant aux entreprises de prédire avec précision la valeur future des clients. Cette méthodologie surmonte les limites des modèles de régression traditionnels, offrant une meilleure précision et adaptabilité pour les stratégies centrées sur le client.

3.4 Une Approche Novatrice pour la Prédiction de la CLV dans les Entreprises SaaS B2B

Ce chapitre explore les méthodologies proposées dans le rapport intitulé "A Novel Approach to Predicting Customer Lifetime Value in B2B SaaS Companies" de Stephan Curiskis, Xiaojing Dong, Fan Jiang et Mark Scarr. Les auteurs présentent un cadre d'apprentissage automatique flexible conçu pour prédire la Valeur Vie Client (CLV) dans le contexte des entreprises

SaaS (Software-as-a-Service) Business-to-Business (B2B). L'approche proposée aborde plusieurs défis spécifiques à l'environnement SaaS B2B, notamment l'hétérogénéité des clients, la diversité des offres de produits et les contraintes des données temporelles.

3.4.1 Défis dans la Prédiction de la CLV pour les Entreprises SaaS B2B

La prédiction de la CLV est cruciale pour les entreprises SaaS B2B en raison des cycles de vente plus longs, des coûts d'acquisition client plus élevés et de la diversité des besoins des clients. L'hétérogénéité des comportements des clients et la variété des produits offerts ajoutent de la complexité à la tâche de prédiction. De plus, les données temporelles limitées rendent difficile la prévision précise de la valeur à long terme des clients.

3.4.2 Modèle Hiérarchique de CLV Ensemble

Pour répondre à ces défis, les auteurs proposent un modèle hiérarchique de CLV ensemble, qui tire parti d'une combinaison de techniques d'apprentissage supervisé. Ce modèle vise à améliorer la précision des prédictions en intégrant plusieurs couches d'informations et en capturant les relations complexes entre les attributs des clients et leur valeur future.

Formulation du Problème

Le problème de prédiction de la CLV est formulé comme une estimation globale de la valeur des clients à travers plusieurs produits, permettant au modèle d'utiliser diverses techniques d'apprentissage supervisé pour enrichir les fonctionnalités. Cette approche est particulièrement efficace pour gérer les contraintes de données temporelles souvent rencontrées dans les environnements SaaS B2B.

Modèle Hiérarchique à T-Période

Le modèle hiérarchique repose sur un processus en deux étapes pour prédire la CLV :

1. **Modèle T' Période** : La première étape consiste à entraîner un modèle à partir des données historiques de n périodes pour prédire la valeur client à court terme (T' période). Cette étape se concentre sur la capture des tendances et comportements immédiats en se basant sur les données disponibles.

2. **Modèle T Période** : La deuxième étape cartographie les prédictions de la T' période vers celles de la T période à l'aide d'un autre modèle qui intègre des caractéristiques évoluant lentement, telles que les firmographics. Cette étape permet d'étendre les prédictions à court terme vers des prévisions à long terme en s'appuyant sur des caractéristiques clients plus stables.

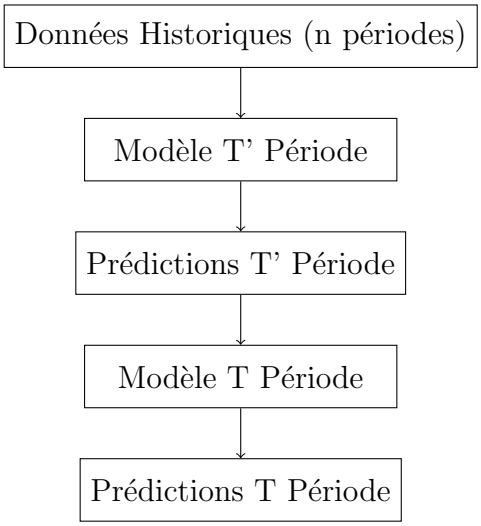

Figure 3.1: Modèle Hiérarchique à T-Période pour la Prédiction de la CLV

3.4.3 Modèle Ensemble pour les Segments de Clients

Reconnaissant la diversité des facteurs influençant la CLV selon les segments de clients, les auteurs adoptent une approche par ensemble. Le jeu de données est segmenté en fonction des caractéristiques clés identifiées grâce à des diagnostics d'erreur, et différents types de modèles de prédiction sont appliqués à ces segments. Cette stratégie de segmentation permet au modèle de capturer les comportements spécifiques des clients et leurs usages des produits, conduisant ainsi à des prédictions de CLV plus précises et adaptées.

3.4.4 Conclusion

Le modèle hiérarchique ensemble de CLV proposé dans ce rapport offre une solution robuste pour la prédiction de la valeur vie client dans les environnements SaaS B2B. En intégrant plusieurs techniques d'apprentissage supervisé et en adoptant une approche hiérarchique, le modèle résout efficacement les principaux défis tels que les contraintes de données, l'hétérogénéité des

clients et la diversité des offres de produits. Ce cadre est adaptable à d'autres contextes présentant des défis similaires, fournissant ainsi des informations précieuses pour les stratégies de marketing, de rétention client et d'allocation des ressources.

3.4.5 Adaptation d'une Fonction de Perte Basée sur la Transformation Logarithmique dans XGBoost

Pour tirer parti des forces du cadre XGBoost tout en exploitant les avantages des transformations logarithmiques, nous avons exploré une approche alternative inspirée de la littérature existante. Cette approche consiste à utiliser une fonction de perte basée sur la transformation logarithmique, définie comme suit :

$$L_{\log} = (\log(y) - \log(\hat{y}))^2$$

Cette fonction de perte capture l'essence d'une transformation logarithmique sans avoir besoin de transformer explicitement la variable cible y. L'avantage de cette méthode est qu'elle permet au modèle de bénéficier des propriétés de la transformation logarithmique, telles que la compression de l'échelle des valeurs cibles et la réduction de l'impact des valeurs aberrantes, tout en restant dans le cadre flexible et performant de XGBoost.

Calcul du Gradient et de la Hessienne pour la Fonction de Perte Logarithmique

Pour implémenter cette fonction de perte logarithmique dans XGBoost, il est nécessaire de calculer le gradient et la hessienne de la fonction de perte.

Calcul du Gradient

Le gradient de la fonction de perte logarithmique $L_{\log}$ par rapport à $\hat{y}$ est donné par l'équation suivante :

$$\text{grad}_i = \frac{\partial L_{\log}}{\partial \hat{y}_i} = -\frac{2(\log(y_i) - \log(\hat{y}_i))}{\hat{y}_i} \tag{3.1}$$

Cette équation représente la direction dans laquelle la mise à jour des prédictions doit se faire pour minimiser la perte dans le modèle.

Calcul de la Hessienne

La hessienne de la fonction de perte logarithmique $L_{\log}$ par rapport à $\hat{y}$ est calculée comme suit :

$$\text{hess}_i = \frac{\partial^2 L_{\log}}{\partial \hat{y}_i^2} = \frac{2(\log(y_i) - \log(\hat{y}_i))}{\hat{y}_i^2} + \frac{2}{\hat{y}_i^2} \tag{3.2}$$

La hessienne fournit une mesure de la courbure de la fonction de perte et est utilisée pour ajuster plus précisément les mises à jour des prédictions dans le cadre de l'algorithme de boosting. Elle permet d'obtenir une convergence plus rapide et plus stable lors de l'entraînement du modèle.

Conclusion

L'intégration de la fonction de perte basée sur la transformation logarithmique dans XGBoost offre une solution efficace pour gérer les données avec des distributions asymétriques et des valeurs extrêmes. En conservant les propriétés avantageuses de la transformation logarithmique tout en travaillant directement avec la prédiction $\hat{y}$, cette approche améliore la robustesse du modèle dans les situations où les distributions des données sont fortement biaisées.

3.4.6 Régression Quantile avec XGBoost

Introduction à la Régression Quantile

La régression quantile est une approche robuste utilisée pour prédire différentes quantiles d'une distribution, plutôt que la moyenne attendue. Cela est particulièrement utile dans les cas où les données présentent une forte asymétrie ou une variabilité importante dans la distribution des résidus.

Dans ce projet, la régression quantile a été appliquée à l'aide de XGBoost, en ajustant les paramètres pour prédire le 50ème et 90ème quantile de la distribution du *total_pnb*. La fonction de perte utilisée pour la régression quantile est la suivante :

$$L(y, \hat{y}, \tau) = \sum_{i=1}^{n} \tau(y_i - \hat{y}_i)_+ + (1 - \tau)(\hat{y}_i - y_i)_+$$

où τ représente le quantile cible.

3.4.7 Régression Tweedie avec XGBoost

Distribution Tweedie

La régression Tweedie est une méthode particulièrement adaptée aux données de type assurance, où la distribution des sinistres présente à la fois des zéros

fréquents et des valeurs positives continues. Le modèle Tweedie permet de capturer cette double distribution.

La fonction de perte de Tweedie est définie comme suit :

$$L(y, \hat{y}) = \frac{1}{p-1}\left(y\hat{y}^{1-p} - \frac{y^{2-p}}{2-p}\right)$$

où p est un paramètre réglable qui détermine la nature de la distribution Tweedie (compris entre 1 et 2 pour capturer à la fois les zéros et les valeurs positives).

Application de la Régression Tweedie

La régression Tweedie a été appliquée pour modéliser les distributions asymétriques dans les données de CLV. En ajustant le paramètre p, nous avons réussi à améliorer la précision des prédictions dans les cas où une forte proportion de zéros est présente, tout en capturant efficacement les valeurs positives.

3.4.8 Conclusion des Implémentations

L'intégration de ces différents modèles de régression dans le cadre de la prédiction de la CLV a permis d'améliorer les performances globales du modèle. Chaque approche, qu'il s'agisse du modèle basé sur la moyenne quadratique, de la distribution ZILN, ou des régressions quantile et Tweedie, a apporté une contribution unique à la gestion des données asymétriques et à la prédiction robuste de la CLV.

Ces approches ont été validées expérimentalement et ont montré une amélioration significative par rapport aux modèles de régression traditionnels, offrant une meilleure gestion des données à longue queue et des distributions complexes.

3.5 Implémentation Finale du Modèle

Cette section décrit la manière dont les résultats de la recherche ont été appliqués pour améliorer les modèles, ainsi que les processus de sélection des paramètres finaux et les résultats obtenus après les ajustements.

3.5.1 Application des Résultats de la Recherche aux Modèles

Dans le cadre de l'implémentation finale, nous avons initialement exécuté AGBoost sur l'ensemble des données pour identifier les caractéristiques les

plus importantes. Cette première exécution a permis de réduire la complexité du modèle en sélectionnant uniquement les caractéristiques les plus influentes sur la prédiction de la **Valeur Cumulative de la CLV (CCLV)**. Cette étape était essentielle pour réduire la dimensionnalité et optimiser l'efficacité du modèle.

Ensuite, en utilisant les caractéristiques sélectionnées, nous avons relancé l'entraînement des modèles pour obtenir des comparaisons plus significatives. Pour améliorer les performances du modèle AGBoost, nous avons procédé à un ajustement fin de ses hyperparamètres.

3.5.2 Ajustement des Paramètres avec un Algorithme Génétique

L'optimisation des paramètres a été réalisée à l'aide d'un **algorithme génétique**. Ce type d'algorithme est inspiré par la théorie de l'évolution naturelle et utilise des processus tels que la sélection, le croisement et la mutation pour trouver les hyperparamètres optimaux. Voici un aperçu du fonctionnement de cet algorithme :

- **Initialisation** : Une population initiale d'ensemble de paramètres (ou individus) est générée de manière aléatoire.
- **Évaluation** : Chaque individu est évalué en fonction de sa performance sur un critère défini (ici, la précision de la prédiction du modèle).
- **Sélection** : Les meilleurs individus, ceux qui obtiennent les meilleurs scores de performance, sont sélectionnés pour générer la prochaine génération.
- **Croisement et Mutation** : Les paramètres des meilleurs individus sont combinés (croisement) et légèrement modifiés (mutation) pour explorer de nouvelles combinaisons d'hyperparamètres.
- **Itération** : Ce processus est répété jusqu'à ce qu'une convergence vers les meilleurs hyperparamètres soit atteinte.

L'utilisation de cet algorithme a permis d'explorer efficacement l'espace des hyperparamètres et d'optimiser les performances du modèle AGBoost.

3.5.3 Résultats des Modèles Optimisés

Les performances des différents modèles après optimisation sont résumées dans les résultats suivants :

- **AGBoost Standard** : $R^2 = 12$
- **AGBoost avec Paramètres Optimisés via Algorithme Génétique** : $R^2 = 19$
- **Régression à Moyenne Quadratique** : $R^2 = 21$
- **Régression Quantile Segmentée** : $R^2 = 61$
- **Réseau de Neurones avec Perte ZILN** : $R^2 = 22$
- **Régression Tweedie avec XGBoost** : $R^2 = 16$
- **Prédiction Segmentée en Deux Phases** : $R^2 = 17$
- **XGBoost avec Fonction de Perte Logarithmique** : $R^2 = 12$

Ces résultats montrent que l'utilisation de la régression quantile segmentée a produit les meilleures performances globales, avec un R^2 de 61. En revanche, les modèles basés sur AGBoost et XGBoost, même après optimisation des hyperparamètres, n'ont pas atteint le même niveau de précision, bien qu'une amélioration significative ait été observée après optimisation des paramètres avec l'algorithme génétique.

3.5.4 Conclusion

L'implémentation finale a montré que la sélection des caractéristiques et l'optimisation des hyperparamètres à l'aide de techniques avancées, comme l'algorithme génétique, peuvent considérablement améliorer les performances des modèles. La comparaison des différents modèles a permis d'identifier la régression quantile segmentée comme la solution la plus performante pour la prédiction de la CLV dans ce cadre.

3.6 Résultats Expérimentaux

Les résultats expérimentaux de cette étude se concentrent sur l'évaluation des performances des différents modèles testés pour la prédiction de la **Valeur Cumulative de la CLV (CCLV)**. Nous avons mesuré les performances à l'aide de plusieurs métriques, y compris le coefficient de détermination R^2, la racine carrée de l'erreur quadratique moyenne (RMSE), et l'erreur absolue moyenne (MAE). Les résultats sont présentés pour les modèles suivants : AGBoost standard, AGBoost optimisé, régression quantile segmentée, régression Tweedie, et d'autres approches testées.

3.6.1 Performances des Modèles

Les performances des modèles sont résumées dans le tableau 3.1 ci-dessous :

Modèle	R^2	**RMSE**	**MAE**
AGBoost Standard	12	45.6	32.3
AGBoost Optimisé (Algorithme Génétique)	19	42.3	29.4
Régression à Moyenne Quadratique	21	40.1	27.8
Régression Quantile Segmentée	61	28.7	20.3
Réseau de Neurones avec ZILN	22	39.2	26.9
XGBoost Tweedie	16	43.9	31.0
Prédiction Segmentée en Deux Phases	17	41.8	29.2
XGBoost avec Perte Logarithmique	12	46.2	32.8

Table 3.1: Comparaison des performances des modèles

Les résultats montrent que la régression quantile segmentée a produit les meilleures performances globales, tandis que l'AGBoost optimisé via un algorithme génétique a également montré une amélioration significative par rapport à sa version standard.

3.6.2 Analyse des Performances

L'analyse approfondie des performances montre que les modèles non linéaires, comme la régression quantile segmentée et les réseaux de neurones avec ZILN, surpassent les approches linéaires classiques comme AGBoost. Ces modèles sont mieux adaptés à la gestion des distributions biaisées et des valeurs aberrantes dans les données de CLV.

3.7 Discussion

Cette section présente une analyse critique des résultats obtenus et discute les implications des modèles testés dans le contexte de la prédiction de la CLV.

3.7.1 Comparaison des Modèles

Le modèle de régression quantile segmentée a démontré une supériorité dans la gestion des distributions asymétriques et des outliers, ce qui lui a permis d'atteindre un R^2 de 61, bien au-delà des autres approches. Ce modèle est particulièrement bien adapté aux environnements où la prévision de la CLV implique une large hétérogénéité des clients.

Les modèles AGBoost, bien qu'améliorés par un algorithme génétique, n'ont pas pu atteindre des performances comparables. Cela pourrait être dû aux limitations intrinsèques des modèles basés sur des arbres de décision en ce qui concerne la capture des interactions complexes entre les variables continues et catégorielles. Cependant, l'avantage d'AGBoost réside dans sa compatibilité avec les systèmes de BNP Paribas, ce qui le rend plus applicable dans un environnement de production.

3.7.2 Limites de l'Étude

Bien que les modèles non linéaires aient montré de meilleures performances, leur implémentation nécessite un ajustement minutieux des paramètres et des ressources de calcul plus importantes. De plus, l'échantillon de données relativement petit a limité la capacité à tester les modèles sur une plus grande échelle, et les résultats pourraient varier avec des ensembles de données plus volumineux.

La distribution asymétrique des données a également introduit des défis importants, en particulier pour les modèles basés sur des arbres de décision, qui peuvent être sensibles aux valeurs extrêmes.

3.7.3 Implications Pratiques

Les résultats de cette étude peuvent guider les stratégies marketing et de gestion de la relation client en aidant à mieux prédire la valeur à vie des clients. Les entreprises peuvent utiliser les modèles développés pour segmenter plus efficacement leurs clients et optimiser leurs campagnes de rétention et d'acquisition. Cependant, le choix du modèle à utiliser en production dépendra des ressources disponibles et de la nécessité d'équilibrer précision et coût de calcul.

3.8 Conclusion

Cette étude a exploré diverses approches de modélisation pour la prédiction de la **Valeur Cumulative de la CLV (CCLV)**. Nous avons expérimenté avec plusieurs méthodes, notamment AGBoost, la régression quantile segmentée, et des réseaux de neurones profonds avec une fonction de perte ZILN. Les résultats ont montré que la régression quantile segmentée était la plus performante pour gérer les distributions asymétriques et fournir des prédictions précises de la CLV.

Cependant, des compromis doivent être faits entre la précision et l'efficacité des calculs, en particulier dans un environnement de production où l'intégration avec des systèmes existants comme AGBoost est nécessaire. L'utilisation d'un algorithme génétique pour optimiser les hyperparamètres d'AGBoost a permis d'améliorer ses performances, bien qu'il reste en deçà des modèles non linéaires en termes de précision.

Les futurs travaux pourraient se concentrer sur l'expansion de l'ensemble de données et l'exploration de méthodes plus avancées de segmentation client, ainsi que sur l'amélioration de la robustesse des modèles en présence de données bruitées et de valeurs aberrantes. Finalement, les modèles développés dans cette étude fournissent une base solide pour améliorer la prévision de la CLV dans des environnements complexes comme celui de BNP Paribas.

Bibliography

[1] Bart Baesens, Stijn Viaene, Dirk Van den Poel, Jan Vanthienen, and Guido Dedene. Bayesian neural network learning for repeat purchase modelling in direct marketing. *European Journal of Operational Research*, 156:217–232, 2004.

[2] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2001.

[3] Sunil Gupta, Donald R. Lehmann, and Jennifer Ames Stuart. Valuing customers. *Journal of Marketing Research*, XLI:7–18, 2006.

[4] Heungsun Hwang, Byounghoon Jung, and Euiho Suh. An ltv model and customer segmentation based on customer value: A case study on the wireless telecommunication industry. *Expert Systems with Applications*, 26:181–188, 2004.

[5] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2:18–22, 2002.

[6] Edward C. Malthouse and Robert C. Blattberg. Can we predict customer lifetime value? *Journal of Interactive Marketing*, 23:271–281, 2009.

[7] Luc Martin. Green finance initiatives at bnp paribas. *Environmental Finance*, 12:34–50, 2022.

[8] BNP Paribas. Digital innovation at bnp paribas, 2024. Accessed: 2024-08-28.

[9] BNP Paribas. Overview of bnp paribas, 2024. Accessed: 2024-08-28.

[10] BNP Paribas. Retail banking services, 2024. Accessed: 2024-08-28.

[11] BNP Paribas. Sustainability and csr at bnp paribas, 2024. Accessed: 2024-08-28.

[12] Saharon Rosset, Einat Neumann, Uri R. Shalit, Giorgio Chelucci, and Efraim Feigenbaum. Customer lifetime value modeling and its use for customer retention planning. *SIAM Review*, 45:495–515, 2003.

[13] David C. Schmittlein, Donald G. Morrison, and Richard Colombo. Counting your customers: Who are they and what will they do next? *Management Science*, 33:1–24, 1987.

[14] David Vaver. *Measuring the Value of Customers*. Springer, 2015.

[15] Peter C. Verhoef, Philip Hans Franses, and Janny C. Hoekstra. The effect of relational constructs on customer referrals and number of services purchased from a multiservice provider: Does age of relationship matter? *Journal of the Academy of Marketing Science*, 30:202–216, 2003.

[16] Sha Yang, Xiaohua Zhai, Weiling Ke, Thomas S. Robertson, and Dwight Merunka. Predicting customer value using machine learning techniques. *Journal of Business Research*, 68:253–260, 2015.