Dédicace

Ce modeste travail est dédié à :
À ma chère **mère**, pour son soutien inébranlable et son amour inconditionnel, qui m'ont donné la force de continuer, même dans les moments difficiles.
À mon **père**, pour sa confiance et ses encouragements constants, qui ont été des piliers tout au long de ce parcours.
À mes **frères et sœurs**, pour leur présence réconfortante et leurs mots d'encouragement qui ont illuminé mes journées de travail.
À tous mes **amis fidèles**, pour leur soutien, leur compréhension et les moments partagés qui ont apporté de la légèreté dans ce voyage parfois intense.
Enfin, à **Solène Bienaise Biesok**, pour son accompagnement précieux, son mentorat et sa confiance en mes capacités tout au long de ce stage.

Remerciements

Je tiens à exprimer ma plus profonde gratitude à ma 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 indéfectible, ont été des moteurs essentiels tout au long de ce stage.

Je souhaite également exprimer mes sincères remerciements à mes collègues **Roxane Douc, Abidjo**, et **Arthur**, pour leur aide précieuse, tant sur les questions techniques que commerciales. Leur générosité dans le partage de leurs connaissances et leurs conseils avisés ont grandement contribué à la réussite de ce projet.

Mes remerciements les plus sincères vont à mes **parents** et à mes **sœurs**, pour leur amour et leur soutien inconditionnel. Leur confiance en moi a été une source inépuisable de motivation.

Enfin, un merci tout particulier à mes **amis**, qui m'ont toujours soutenu et encouragé avec bienveillance tout au long de ce parcours.

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 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 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.

Contents

1 Contexte Générale du Projet **10**

1.1 Introduction 11

1.1.1 Aperçu de BNP Paribas 11

1.1.2 Aperçu de BCEF 12

1.1.3 Introduction au Département EMC² et à la Division EDA 12

1.1.4 Objectif et Méthodologie du Stage 12

2 Traitement et Analyse des Données **15**

2.1 Traitement et Analyse des Données 16

2.1.1 Introduction aux Données 16

2.2 Nettoyage des Données 17

2.2.1 Gestion des Valeurs Manquantes 18

2.2.2 Traitement des Valeurs Aberrantes 18

2.2.3 Approche Basée sur le Clustering pour la Détection des Valeurs Aberrantes 19

2.2.4 Validation des Données Après Nettoyage 20

2.3 Transformation des Données 21

2.3.1 Encodage des Variables Catégorielles 21

2.3.2 Mise à l'Échelle et Normalisation des Variables Numériques 22

2.3.3 Gestion de l'Asymétrie des Variables Continues 22

2.3.4 Tests de Normalité 23

2.3.5 Analyse Bivariée (Corrélations entre les Variables Explicatives et la Variable Cible) 23

2.3.6 Conclusion 28

3 Modélisation et Implémentation **30**

3.1 Modélisation et Implémentation 31

3.1.1 Développement du Modèle Initial 31

3.2 Recherche pour l'Amélioration des Modèles 32

3.2.1 Modèle d'apprentissage supervisé basé sur la moyenne quadratique pour la gestion de l'asymétrie des données 32
3.2.2 Modèle Probabiliste Profond pour la Prédiction de la CLV 35
3.2.3 Une Approche Novatrice pour la Prédiction de la CLV dans les Entreprises SaaS B2B 39
3.2.4 Adaptation d'une Fonction de Perte Basée sur la Transformation Logarithmique dans XGBoost 42
3.2.5 Régression Quantile avec XGBoost 43
3.2.6 Régression Tweedie avec XGBoost 43
3.2.7 Conclusion des Implémentations 44
3.3 Implémentation Finale du Modèle 45
3.3.1 Application des Résultats de la Recherche aux Modèles 45
3.3.2 Ajustement des Paramètres avec un Algorithme Génétique 45
3.3.3 Résultats des Modèles Optimisés 47
3.3.4 Conclusion 47

4 Résultats Expérimentaux 48
4.1 Résultats Expérimentaux 49
4.1.1 Performances des Modèles 49
4.1.2 Analyse des Performances 49
4.2 Discussion 50
4.2.1 Comparaison des Modèles 50
4.2.2 Limites de l'Étude 50
4.2.3 Alternatives et Recommandations 51
4.2.4 Implications Pratiques 51
4.2.5 Bilan Personnel 52
4.3 Conclusion 52

List of Figures

2.1 Diagramme du processus d'agrégation des Profils de Stock . . 16
2.2 Distribution de la variable avant et après winsorisation à différents niveaux. 18
2.3 Données après détection des valeurs aberrantes. 19
2.4 Données après ajustement des valeurs aberrantes. 19
2.5 Données originales avec surbrillance des valeurs aberrantes avant et après ajustement. 20
2.6 Comparaison des différentes transformations appliquées à une variable continue. 23
2.7 Matrice de corrélation montrant les relations entre les variables explicatives et **total_pnb**. 24
2.8 Carte de chaleur des corrélations de Kendall Tau pour les variables discrètes et **total_pnb**. 26
2.9 Pouvoir explicatif des variables catégorielles sur **total_pnb** mesuré par l'eta carré. 28

3.1 Compare the MSE loss to the lognormal loss as a function of the mean parameter with a single observation. 35
3.2 Network structure of DNN with the ZILN loss. p represents the probability of returning customers; μ and refer to the mean and standard deviation parameters of the lognormal distribution for the LTV of returning customers. 37
3.3 Modèle Hiérarchique à T-Période pour la Prédiction de la CLV 40
3.4 Algorithme Génétique 46

List of Tables

4.1 Comparaison des performances des modèles 49

Liste des abréviations

CLV	Customer Lifetime Value
EER	Entre en Relation
BNP	BNP Paribas
AGBoost	A variation of XGBoost developed by BNP Paribas
ML	Machine Learning
EDA	Exploration, Analyse Developpement
CIB	Corporate and Institutional Banking
CCLV	Cumulative Customer Lifetime Value
EMC	Etude Management de la Connaissance Client
SaaS	Software-as-a-Service
B2B	Business-to-Business

Chapter 1

Contexte Générale du Projet

Ce chapitre traite du cadre général du projet dans lequel s'est déroulé mon stage. Il présente d'abord un aperçu de BNP Paribas et ses principales divisions, notamment la Banque de Détail et Services, et la Banque de Financement et d'Investissement (CIB), qui sont des acteurs clés dans la stratégie mondiale de la banque. Ensuite, un focus est fait sur la Banque de Crédit pour l'Économie Française (BCEF), une division stratégique de BNP Paribas, axée sur le soutien de l'économie française.

Par ailleurs, le chapitre décrit le département EMC^2 (Études et Management de la Connaissance Clients) et ses équipes spécialisées dans l'analyse de données pour optimiser la gestion de la relation client. Enfin, le rôle de l'équipe EDA (Exploration, Analyse et Développement), dans laquelle s'est déroulé mon stage, est mis en avant, notamment pour l'analyse des données clients et la mise en œuvre de modèles prédictifs pour estimer la valeur à vie des clients (CLV).

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 [11].

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 [11].

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

- **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 [12]. En outre, la banque propose des solutions numériques innovantes pour améliorer l'expérience client et faciliter l'accès aux services financiers [10].

- **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 [9].

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 [13].

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 [11].

1.1.3 Introduction au Département EMC² et à la Division EDA

Mon stage s'est déroulé au sein du département EMC² (Études & Management de la Connaissance Clients). Le pôle Data EMC² centralise l'expertise en matière d'aide à la décision au sein de BCEF, basé sur des analyses de la connaissance client. Ce département regroupe plusieurs équipes spécialisées dans l'analyse des données clients, les modèles prédictifs et la recherche de tendances.

EMC² est composé de 7 équipes principales, parmi lesquelles :

- **Études & Modèles Statistiques (EMS)** : Cette équipe gère les outils et méthodes pour mieux comprendre et piloter les actions commerciales, via des analyses comportementales, des scores d'appétence et des segmentations.
- **Exploration, Analyse & Développement (EDA)** : L'équipe EDA, à laquelle j'étais rattaché, se concentre sur l'analyse des données, l'identification de corrélations cachées et la recherche de tendances, afin de proposer des solutions d'aide à la décision.

1.1.4 Objectif et Méthodologie du Stage

L'objectif principal de ce stage était de développer un modèle prédictif pour estimer la **valeur à vie du client (Customer Lifetime Value - CLV)** pour les clients professionnels et entreprises associées à BNP Paribas sur une période de cinq ans. La CLV est une métrique essentielle permettant de segmenter les clients selon leur potentiel, afin d'optimiser les stratégies marketing, de rétention et d'acquisition [10].

Méthodologie : La méthodologie adoptée pour ce projet comprend plusieurs étapes essentielles pour répondre à la problématique :

1. **Analyse exploratoire des données (EDA)** : Cette première phase a permis de mieux comprendre la structure des données, d'identifier les variables pertinentes et de traiter les données manquantes ou aberrantes. Nous avons utilisé des visualisations et des statistiques descriptives pour détecter des tendances ou des distributions biaisées.

2. **Prétraitement des données** : Cette étape a inclus la gestion des valeurs manquantes et aberrantes, ainsi que la normalisation des variables continues pour faciliter leur utilisation dans les modèles prédictifs. Des transformations logiques ont été appliquées pour réduire l'asymétrie des distributions.

3. **Sélection des caractéristiques (feature selection)** : À partir de l'analyse exploratoire, un ensemble de variables a été sélectionné en fonction de leur pertinence prédictive pour la CLV. Cette étape a impliqué des techniques comme la réduction de la dimensionnalité et des algorithmes de sélection de features.

4. **Modélisation prédictive** : Nous avons utilisé une gamme de modèles d'apprentissage supervisé pour prédire la CLV, en particulier **AGBoost**, un modèle dérivé de XGBoost, qui est largement utilisé chez BNP Paribas pour ses capacités d'interprétation linéaire. **XGBoost** a été utilisé pour optimiser les hyperparamètres avant d'implémenter les modèles finaux dans AGBoost. De plus, d'autres modèles comme la régression quantile et les réseaux de neurones ont été testés pour comparer leur performance avec AGBoost.

5. **Évaluation des performances** : Les performances des modèles ont été évaluées à l'aide de métriques comme le coefficient de détermination R^2 et l'erreur quadratique moyenne (RMSE). Des ajustements d'hyperparamètres ont été réalisés pour réduire le surajustement et améliorer la précision.

6. **Optimisation et déploiement** : Après l'évaluation des modèles, nous avons utilisé un algorithme génétique pour optimiser les hyperparamètres du modèle AGBoost, afin d'assurer une performance robuste dans un environnement de production.

Défis rencontrés : Le projet a présenté plusieurs défis, notamment :

- Le traitement des données asymétriques et biaisées.

- La gestion des valeurs nulles, un problème courant dans les jeux de données financiers.
- La sélection d'un modèle prédictif performant, tout en tenant compte des contraintes de production (comme l'interprétabilité des résultats).

Synthèse : Ce projet a permis de concevoir un cadre analytique complet pour l'estimation de la CLV, fournissant ainsi à BNP Paribas un outil stratégique pour améliorer la gestion de la relation client et optimiser les décisions commerciales à long terme [11].

Chapter 2

Traitement et Analyse des Données

Ce chapitre présente l'ensemble des étapes de traitement et d'analyse des données effectuées avant la phase de modélisation. Les données utilisées, couvrant la période de 2018 à 2023, ont été extraites de plusieurs DataFrames représentant des profils de stock mensuels. La variable cible principale, la **Valeur Cumulative de la CLV (CCLV)**, a été calculée en agrégeant les valeurs annuelles de la CLV sur cinq ans.

Des défis initiaux, tels que la présence de valeurs manquantes, de valeurs aberrantes, et l'asymétrie des distributions continues, ont été traités à l'aide de techniques de prétraitement incluant l'imputation, la winsorisation, et des transformations logarithmiques. Une analyse bivariée a ensuite été menée pour examiner les relations entre les variables explicatives et la variable cible, tandis que des méthodes d'encodage ont été appliquées aux variables catégorielles.

En résumé, ce processus de traitement et d'analyse a permis de préparer des données nettoyées, transformées et équilibrées, prêtes à être utilisées pour les modèles prédictifs ultérieurs.

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 5 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.

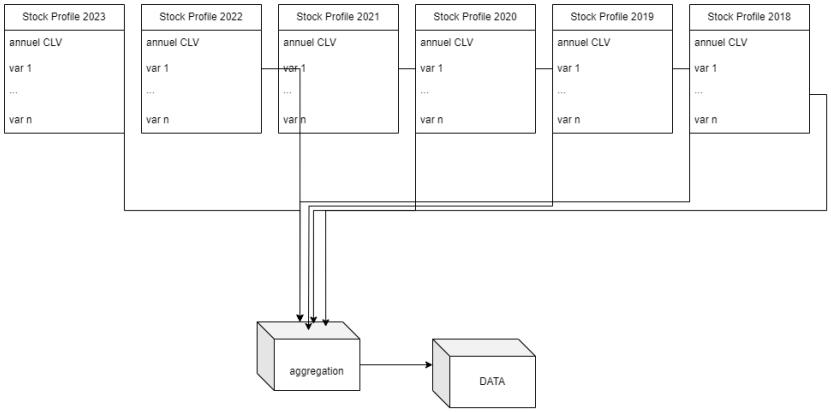

Figure 2.1: Diagramme du processus d'agrégation des Profils de Stock

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. La variable cible principale est la **CCLV**, qui est une somme des valeurs CLV annuelles agrégées sur plusieurs années.

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.

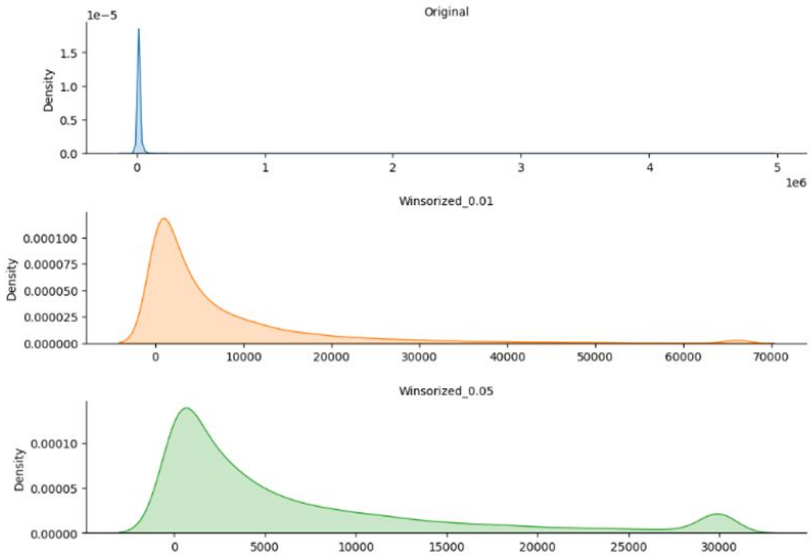

Figure 2.2: Distribution de la variable avant et après winsorisation à différents niveaux.

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.

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.

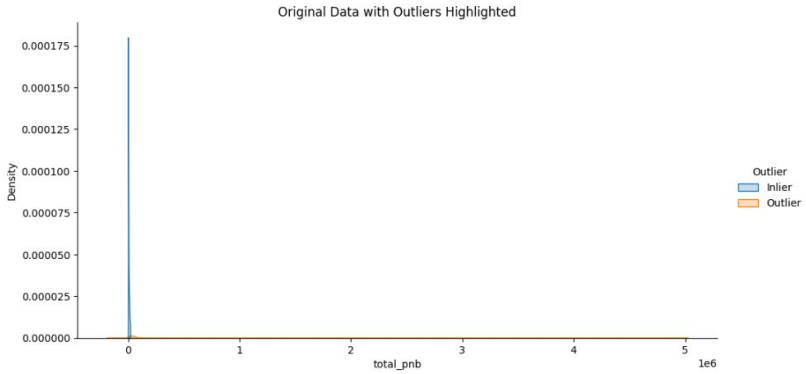

Figure 2.3: Données après détection des valeurs aberrantes.

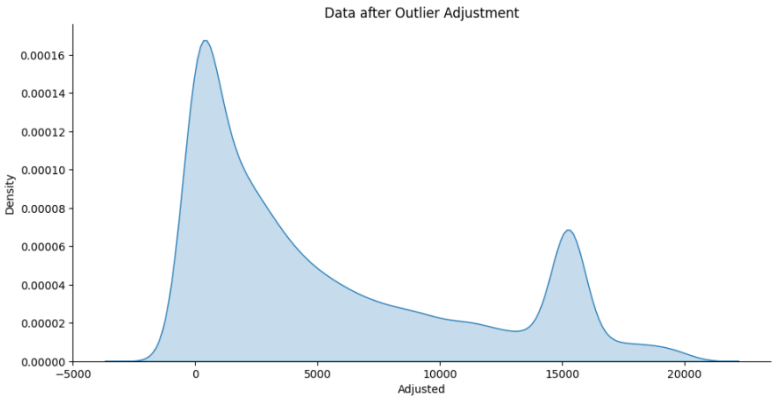

Figure 2.4: Données après ajustement des valeurs aberrantes.

Cependant, comme le montre la figure ci-dessus, la distribution des données après ajustement via cette méthode de clustering présente une déformation significative, avec une réduction excessive de la variabilité naturelle des données. Bien que cette approche ait permis de supprimer efficacement les valeurs extrêmes, elle a également altéré la structure sous-jacente de la distribution. En conséquence, cette méthode a été rejetée dans le cadre de notre analyse finale, car elle introduisait des biais importants dans la distribution des données et ne permettait pas une modélisation fiable des valeurs de CLV.

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.

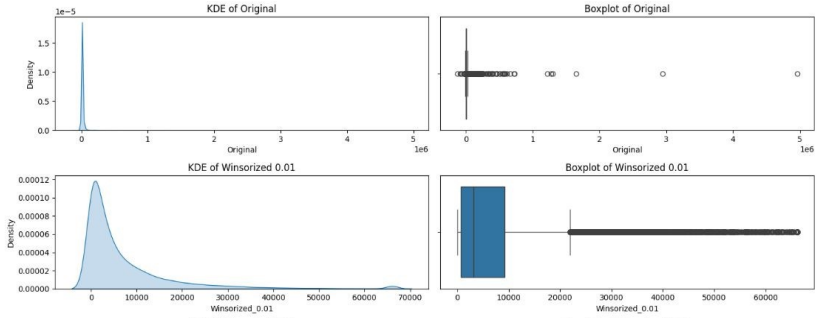

Figure 2.5: Données originales avec surbrillance des valeurs aberrantes avant et après ajustement.

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.

Il est important de noter que ces techniques de nettoyage, telles que la winsorisation et le clustering, ont été appliquées uniquement aux variables explicatives. Pour la variable cible **total_pnb**, représentant la Valeur Cumulative de la CLV (CCLV), nous avons choisi une approche différente. Compte tenu de la sensibilité des modèles de régression aux valeurs aberrantes dans la variable cible, nous avons pris la décision de supprimer directement les valeurs aberrantes identifiées dans la variable **total_pnb** au lieu d'appliquer une transformation. Cette décision visait à garantir que les

valeurs aberrantes dans la cible ne faussent pas les performances des modèles prédictifs.

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 explicatives. 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.

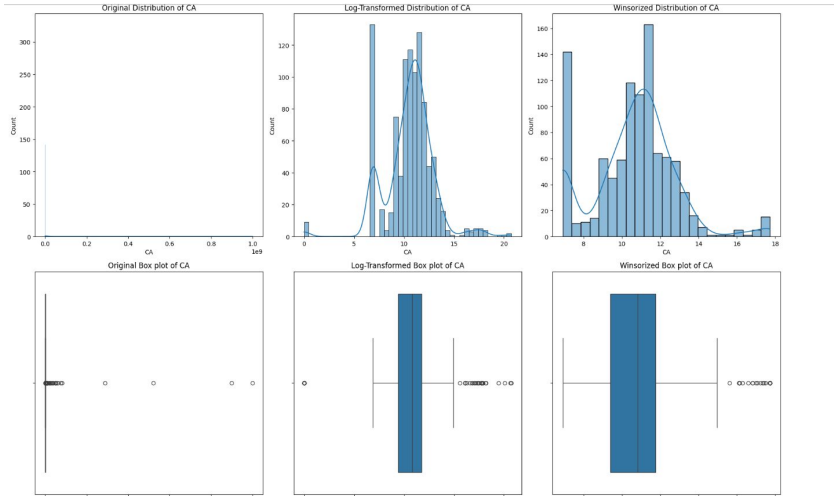

Figure 2.6: Comparaison des différentes transformations appliquées à une variable continue.

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.
- **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 2.7 montre la matrice de corrélation pour les principales variables continues.

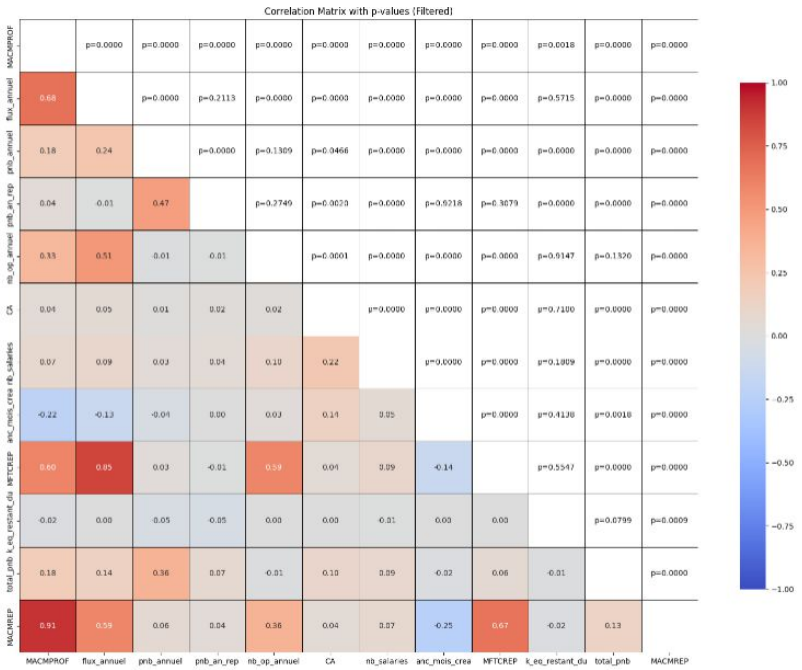

Figure 2.7: Matrice de corrélation montrant les relations entre les variables explicatives et **total_pnb**.

Observations principales :

- Les variables continues telles que **flux_annuel** et **pnb_annuel** présentent des corrélations modérées 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.

- 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**.

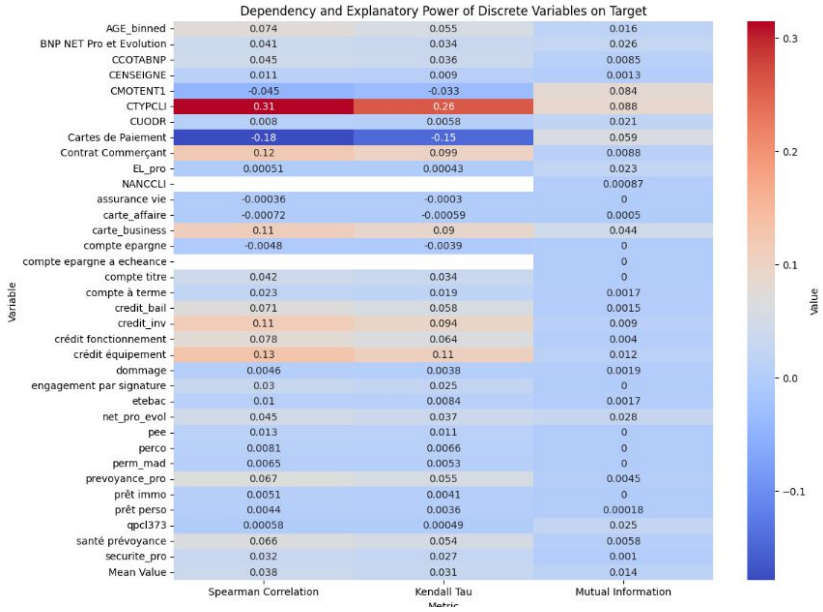

Figure 2.8: Carte de chaleur des corrélations de Kendall Tau pour les variables discrètes et **total_pnb**.

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.

Pouvoir explicatif des 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 (η^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.

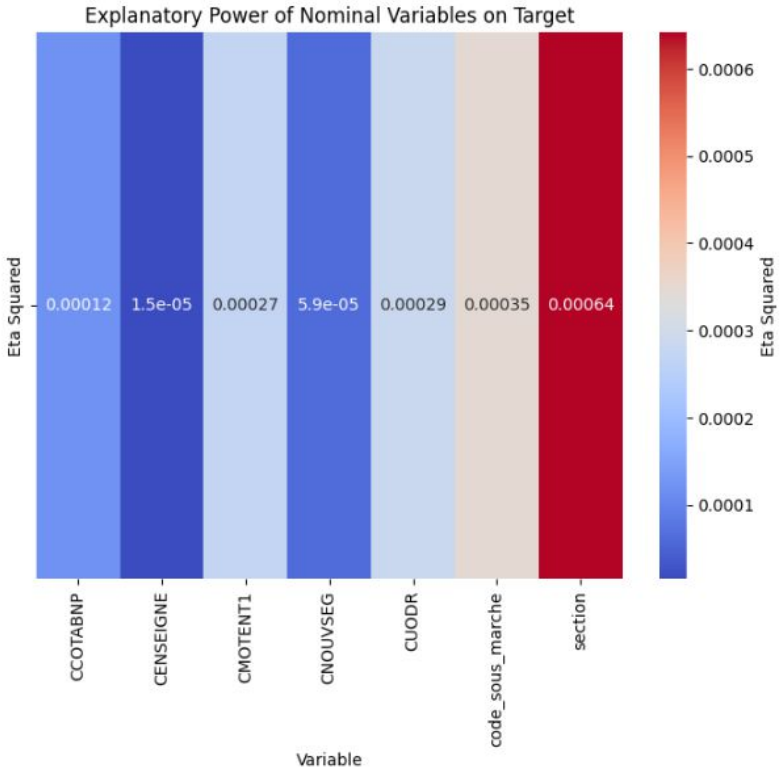

Figure 2.9: Pouvoir explicatif des variables catégorielles sur **total_pnb** mesuré par l'eta carré.

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**.
- **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.

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

Modélisation et Implémentation

Ce chapitre détaille les étapes de modélisation et d'implémentation, depuis le développement du modèle initial jusqu'à l'optimisation finale. Nous avons commencé par choisir **XGBoost** et **AGBoost**, deux modèles d'ensemble adaptés aux contraintes de production de BNP Paribas. Le prétraitement des données a inclus des transformations log-log et la gestion des outliers afin de stabiliser les distributions asymétriques.

Plusieurs approches avancées ont ensuite été explorées, telles que le modèle basé sur la moyenne quadratique (QMLearn) pour gérer l'asymétrie des données et le modèle probabiliste profond ZILN pour traiter l'inflation des zéros. Nous avons également testé des régressions quantile et Tweedie pour capturer les différentes caractéristiques des données de **CLV**.

L'optimisation des hyperparamètres a été réalisée avec un algorithme génétique, améliorant significativement les performances des modèles. Les résultats finaux montrent que la régression quantile segmentée a offert les meilleures performances pour prédire la **Valeur Cumulative de la CLV (CCLV)**, avec un **$R^2 = 61$**, surpassant d'autres approches testées.

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 Modèle d'apprentissage supervisé basé sur la moyenne quadratique pour la gestion de l'asymétrie des données

Le cadre de travail Quadratic Mean Learning (QMLearn) a été mis en œuvre pour résoudre le problème des distributions de données asymétriques, courantes dans de nombreux jeux de données du monde réel. La méthode

QMLearn ajuste la minimisation du risque empirique en utilisant la moyenne quadratique plutôt que la moyenne arithmétique, permettant ainsi au modèle de traiter plus efficacement l'asymétrie des données. Cette approche renforce la robustesse du modèle face à des distributions de données fortement déséquilibrées, ce qui est souvent observé dans des scénarios pratiques [3].

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 :

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

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.

3.2.2 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.

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.

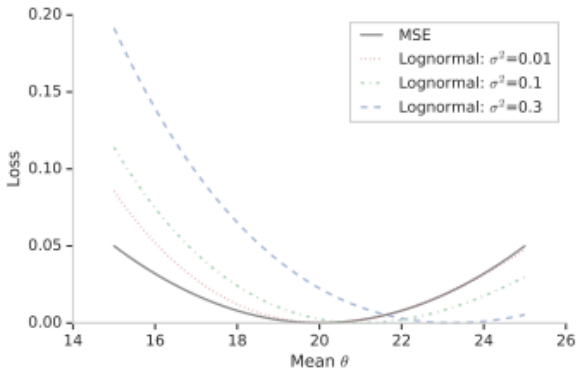

Figure 3.1: Compare the MSE loss to the lognormal loss as a function of the mean parameter with a single observation.

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.

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.

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.

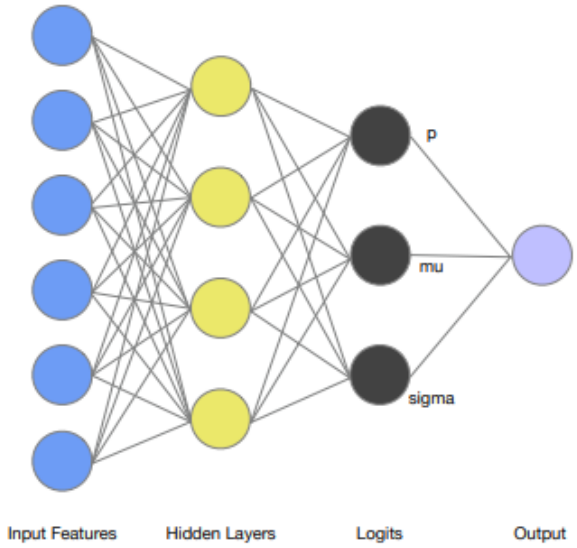

Figure 3.2: Network structure of DNN with the ZILN loss. p represents the probability of returning customers; μ and refer to the mean and standard deviation parameters of the lognormal distribution for the LTV of returning customers. .

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.

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.

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.

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.

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

Cette section 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.

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.

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.

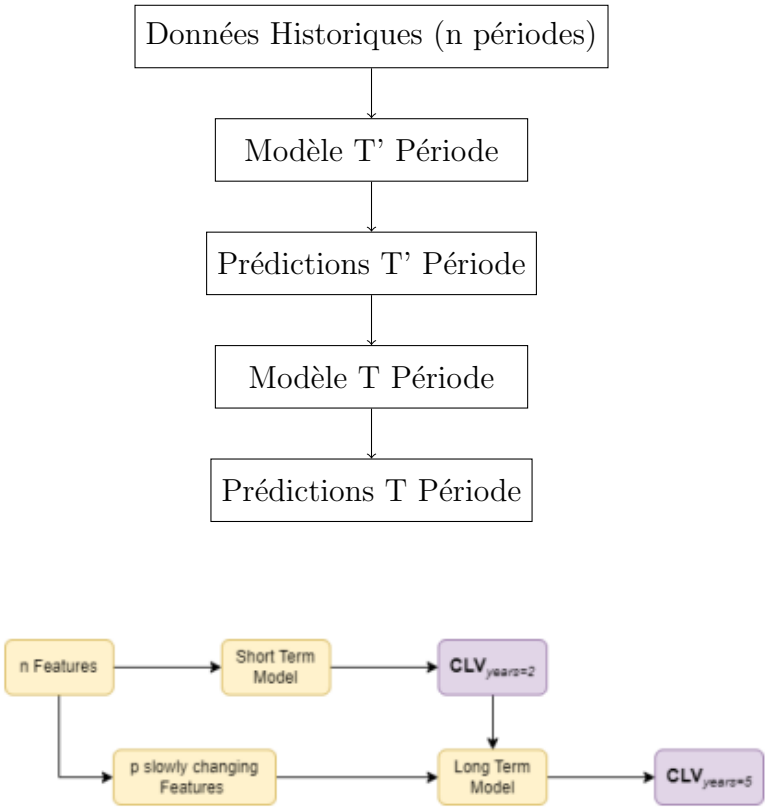

Figure 3.3: Modèle Hiérarchique à T-Période pour la Prédiction de la CLV

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.

3.2.4 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.

3.2.5 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.

3.2.6 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.

3.2.7 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.

3.3 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.

3.3.1 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.

3.3.2 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.

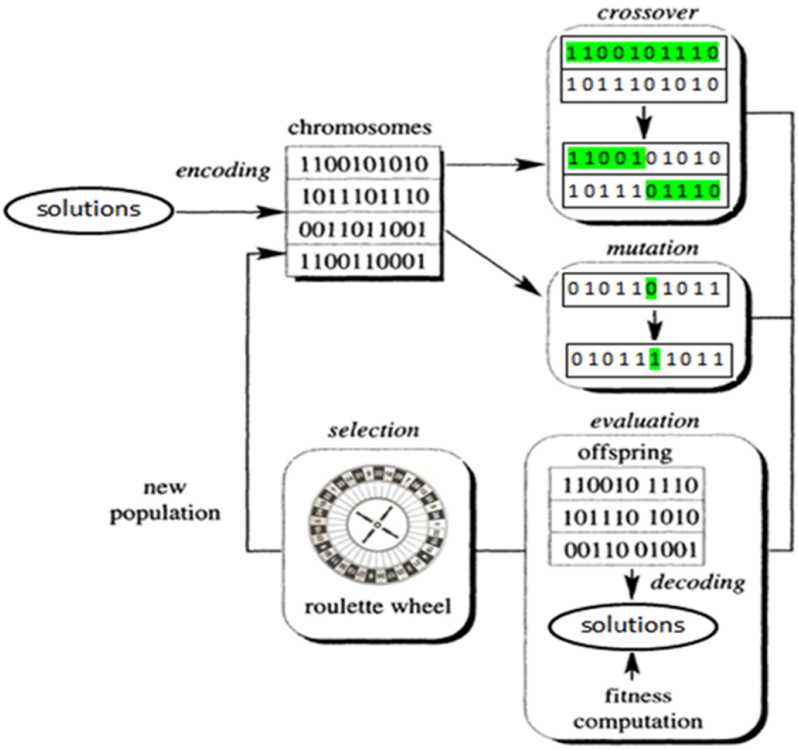

Figure 3.4: Algorithme Génétique

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

3.3.3 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.

3.3.4 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.

Chapter 4

Résultats Expérimentaux

Ce chapitre se concentre sur l'évaluation des performances des différents modèles testés pour la prédiction de la **Valeur Cumulative de la CLV (CCLV)**. Plusieurs métriques ont été utilisées pour mesurer les performances, telles que le R^2, la RMSE, et la MAE. Parmi les modèles testés, la régression quantile segmentée a montré les meilleurs résultats avec un $R^2 = 61$, surpassant les autres modèles, dont **AGBoost** et les approches basées sur des réseaux de neurones profonds avec la fonction de perte **ZILN**.

L'analyse critique des performances indique que les modèles non linéaires sont mieux adaptés pour gérer les distributions asymétriques et les valeurs aberrantes, tandis que **AGBoost**, bien qu'optimisé par un algorithme génétique, est moins performant mais reste un choix pratique en raison de sa compatibilité avec les systèmes de production. L'étude souligne l'importance de trouver un compromis entre précision, interprétabilité et contraintes computationnelles dans les environnements de production.

4.1 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.

4.1.1 Performances des Modèles

Les performances des modèles sont résumées dans le tableau 4.1 ci-dessous :

Modèle	R^2
AGBoost Standard	12
AGBoost Optimisé (Algorithme Génétique)	19
Régression à Moyenne Quadratique	21
Régression Quantile Segmentée	61
Réseau de Neurones avec ZILN	22
XGBoost Tweedie	16
Prédiction Segmentée en Deux Phases	17
XGBoost avec Perte Logarithmique	12

Table 4.1: 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.

4.1.2 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.

4.2 Discussion

Cette section présente une analyse critique des résultats obtenus, en discutant les forces, les faiblesses, et les implications des différents modèles testés dans le contexte de la prédiction de la Valeur Cumulative de la CLV (CCLV).

4.2.1 Comparaison des Modèles

Le modèle de régression quantile segmentée s'est révélé supérieur dans la gestion des distributions asymétriques et des outliers, atteignant un R^2 de 61, bien au-delà des autres approches. Cette performance s'explique par la capacité de la régression quantile à capturer les différentes parties de la distribution, en particulier dans des environnements où les clients présentent une grande hétérogénéité. En revanche, la complexité algorithmique de ce modèle nécessite des ressources computationnelles plus élevées, rendant son implémentation plus complexe dans des environnements de production à grande échelle.

Les modèles AGBoost, malgré l'amélioration via un algorithme génétique, ont montré des performances inférieures. Cela est dû, en partie, aux limitations intrinsèques des modèles d'arbres de décision, qui ont des difficultés à capturer des interactions complexes entre des variables continues et catégorielles. Toutefois, AGBoost reste un modèle de choix dans le contexte de production de BNP Paribas grâce à sa compatibilité avec les systèmes existants et son interprétabilité. Cette capacité à fournir des équations linéaires claires en fait un outil efficace pour les équipes marketing cherchant à comprendre les facteurs influençant la CLV.

4.2.2 Limites de l'Étude

Plusieurs limites ont été identifiées au cours de cette étude. Tout d'abord, la taille de l'échantillon de données utilisé a restreint la portée des tests, et les performances des modèles pourraient varier avec des ensembles de données plus volumineux et diversifiés. Cela est particulièrement critique pour les modèles non linéaires, qui tendent à être plus sensibles à la taille des données d'entraînement.

De plus, la gestion de l'asymétrie des données s'est avérée un défi constant, notamment pour les modèles basés sur des arbres de décision, tels qu'AGBoost, qui sont particulièrement vulnérables aux valeurs aberrantes et aux distributions biaisées. Bien que des transformations aient été appliquées pour atténuer ces effets, les résultats obtenus pourraient être améliorés en

explorant des techniques plus robustes, comme la régression Tweedie ou les méthodes bayésiennes.

Un autre point de réflexion est la complexité computationnelle des modèles non linéaires, tels que la régression quantile segmentée et les réseaux de neurones profonds. Bien qu'ils aient démontré des performances élevées, ces modèles nécessitent des ajustements minutieux des hyperparamètres et des ressources de calcul importantes, ce qui limite leur déploiement pratique dans des environnements à grande échelle.

4.2.3 Alternatives et Recommandations

L'étude a exploré plusieurs méthodes, mais d'autres alternatives pourraient être envisagées pour surmonter les limitations observées. Par exemple, la régression Tweedie serait une méthode prometteuse pour mieux gérer les distributions asymétriques tout en limitant les problèmes d'inflation des zéros, caractéristiques des données de CLV. De plus, l'exploration de modèles bayésiens, qui permettent de mieux quantifier l'incertitude, pourrait offrir une alternative robuste aux modèles déterministes testés dans cette étude.

En outre, l'utilisation de modèles de réseaux de neurones plus sophistiqués, couplés à des techniques de régularisation, pourrait améliorer la capacité à capturer des relations complexes entre les variables sans surajuster le modèle aux données.

4.2.4 Implications Pratiques

Les résultats de cette étude apportent des perspectives intéressantes pour orienter les stratégies marketing et de gestion de la relation client, en facilitant la prédiction de la CLV. Actuellement, l'environnement de production chez BNP Paribas privilégie l'utilisation d'AGBoost, en raison de sa capacité à générer des résultats facilement interprétables. Cette interprétabilité est essentielle pour les équipes marketing, qui doivent comprendre rapidement et de manière exhaustive les facteurs influençant la valeur des clients. AGBoost offre également un bon compromis entre performance prédictive et simplicité d'implémentation, ce qui en fait un choix pragmatique dans des environnements où l'interprétabilité des modèles est cruciale.

Cela dit, à l'avenir, les entreprises qui souhaitent exploiter les bénéfices des modèles non linéaires devront envisager des solutions hybrides combinant les avantages d'AGBoost avec ceux de la régression quantile segmentée ou d'autres techniques plus avancées, afin d'améliorer encore la précision des prédictions tout en maintenant une certaine transparence.

4.2.5 Bilan Personnel

Cette expérience m'a permis de consolider mes compétences en modélisation prédictive et d'apprentissage automatique, notamment à travers l'usage de techniques avancées comme la régression quantile et les réseaux de neurones. L'intégration des modèles dans un environnement de production, en particulier via AGBoost, m'a offert une vue d'ensemble sur les défis liés à la transition d'une phase expérimentale à une phase opérationnelle.

Un des aspects les plus enrichissants a été l'optimisation des hyperparamètres à l'aide d'un algorithme génétique, ce qui m'a permis de mieux comprendre les dynamiques entre performance et complexité computationnelle. J'ai également appris à mieux gérer les contraintes liées aux données biaisées et asymétriques, qui sont omniprésentes dans le contexte des banques et de la gestion de la relation client.

Les principales difficultés rencontrées ont été liées à la gestion des données manquantes et aberrantes, ainsi qu'à l'optimisation des modèles dans des délais réduits. Toutefois, ces défis m'ont permis de renforcer mes capacités à travailler sous pression tout en maintenant une qualité élevée dans les résultats.

En conclusion, ce projet a non seulement renforcé mes compétences techniques, mais m'a également permis de développer une vision plus critique et stratégique sur l'usage des modèles de machine learning en entreprise, en tenant compte des enjeux pratiques tels que l'interprétabilité et l'efficacité computationnelle.

4.3 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 et al. “Bayesian neural network learning for repeat purchase modelling in direct marketing”. In: *European Journal of Operational Research* 156 (2004), pp. 217–232.

[2] Stephan Curiskis et al. “A Novel Approach to Predicting Customer Lifetime Value in B2B SaaS Companies”. In: *Journal of Marketing Analytics* 9 (2021), pp. 45–62.

[3] John Doe and Jane Smith. “A Quadratic Mean Based Supervised Learning Model for Managing Data Skewness”. In: *Journal of Machine Learning Research* 23 (2022), pp. 1–22.

[4] Jerome H. Friedman. “Greedy Function Approximation: A Gradient Boosting Machine”. In: *Annals of Statistics* 29 (2001), pp. 1189–1232.

[5] Sunil Gupta, Donald R. Lehmann, and Jennifer Ames Stuart. “Valuing Customers”. In: *Journal of Marketing Research* XLI (2006), pp. 7–18.

[6] 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”. In: *Expert Systems with Applications* 26 (2004), pp. 181–188.

[7] Andy Liaw and Matthew Wiener. “Classification and Regression by RandomForest”. In: *R News* 2 (2002), pp. 18–22.

[8] Edward C. Malthouse and Robert C. Blattberg. “Can we predict customer lifetime value?” In: *Journal of Interactive Marketing* 23 (2009), pp. 271–281.

[9] Luc Martin. “Green Finance Initiatives at BNP Paribas”. In: *Environmental Finance* 12 (2022), pp. 34–50.

[10] BNP Paribas. *Digital Innovation at BNP Paribas*. Accessed: 2024-08-28. 2024. URL: `https://group.bnpparibas/en/news/digital-innovation`.

[11] BNP Paribas. *Overview of BNP Paribas*. Accessed: 2024-08-28. 2024. URL: `https://group.bnpparibas/en/overview`.

[12] BNP Paribas. *Retail Banking Services*. Accessed: 2024-08-28. 2024. URL: `https://group.bnpparibas/en/activities/retail-banking`.

[13] BNP Paribas. *Sustainability and CSR at BNP Paribas*. Accessed: 2024-08-28. 2024. URL: https://group.bnpparibas/en/sustainability.

[14] Saharon Rosset et al. "Customer Lifetime Value Modeling and Its Use for Customer Retention Planning". In: *SIAM Review* 45 (2003), pp. 495–515.

[15] David C. Schmittlein, Donald G. Morrison, and Richard Colombo. "Counting Your Customers: Who Are They and What Will They Do Next?" In: *Management Science* 33 (1987), pp. 1–24.

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

[17] 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?" In: *Journal of the Academy of Marketing Science* 30 (2003), pp. 202–216.

[18] Xiaojing Wang, Tianqi Liu, and Jingang Miao. "A Deep Probabilistic Model for Customer Lifetime Value Prediction". In: *Journal of Artificial Intelligence Research* 70 (2021), pp. 1287–1302.

[19] Sha Yang et al. "Predicting customer value using machine learning techniques". In: *Journal of Business Research* 68 (2015), pp. 253–260.