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 préliminaires montrent une amélioration significative de la précision des prédictions de CLV, permettant une segmentation plus fine des clients et une optimisation des stratégies de rétention et d'engagement. Ces résultats 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.

Preliminary results indicate a significant improvement in the accuracy of CLV predictions, enabling more precise client segmentation and optimization of retention and engagement strategies. The outcomes of this study are expected to contribute significantly to understanding customer behaviors and optimizing client engagement strategies at BNP Paribas.

Introduction

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].

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].

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].

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].

Importance de la CLV pour BNP Paribas

La modélisation de la CLV est cruciale pour BNP Paribas car elle permet à la banque de mieux comprendre la valeur de ses clients sur le long terme. En optimisant les stratégies d'engagement et de rétention des clients, la banque peut non seulement maximiser ses revenus, mais aussi renforcer la satisfaction et la fidélité de sa clientèle. Ce type d'analyse est particulièrement important dans le secteur bancaire où les relations client-fournisseur peuvent durer plusieurs décennies et où la concurrence est féroce.

Structure du Rapport

Ce rapport est structuré comme suit :

- **Introduction** : Présente BNP Paribas, BCEF, le département EMC2, et la division EDA, ainsi que les objectifs du stage.
- **Revue de la Littérature** : Analyse des concepts clés de la valeur à vie du client (CLV) et des approches d'apprentissage automatique pour la modélisation de la CLV.
- **Traitement et Analyse des Données** : Décrit les étapes de traitement des données, y compris la collecte, le nettoyage, la normalisation, et la gestion des valeurs aberrantes.
- **Modélisation et Mise en Œuvre** : Détaille les modèles d'apprentissage automatique utilisés et leur mise en œuvre pour la modélisation de la CLV.
- **Résultats Expérimentaux** : Présente et analyse les résultats obtenus avec les modèles développés.
- **Discussion** : Interprète les résultats, discute des limites de l'étude, et propose des perspectives de recherche future.
- **Conclusion** : Résume les conclusions principales, propose des recommandations pour BNP Paribas, et inclut une réflexion personnelle sur le stage.

Revue de la Littérature

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.

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.

Analyse Critique de la Littérature

La littérature existante sur la modélisation de la CLV montre un large éventail d'approches, mais il existe encore des lacunes importantes, notamment en ce qui concerne la gestion des données asymétriques et l'inflation des zéros. De plus, bien que de nombreux modèles sophistiqués aient été développés, leur application pratique dans des environnements commerciaux réels reste limitée. Ce stage vise à combler certaines de ces lacunes en appliquant et en adaptant des modèles d'apprentissage automatique à des données bancaires complexes.

Résumé de la Revue de Littérature

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. Toutefois, il est essentiel de continuer à explorer et à adapter ces modèles pour répondre aux besoins spécifiques du secteur bancaire et à la dynamique changeante des marchés.

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.