ROYAUME DU MAROC
UNIVERSITÉ ABDELMALEK ESSAÂDI
FACULTÉ DES SCIENCES ET TECHNIQUES DE TANGER

Projet de fin d'études

En vue de l'obtention du

Diplôme de Master en Modélisation Mathématique et Science de Données

Intitulé

Une Approche Mathématique et d'Intelligence Artificielle pour la Prédiction et la Gestion du Churn Client : Classification, Analyse de Survie et Estimation Probabiliste de la Valeur Vie Client (CLV).

Stage effectué au sein de

Smart Automation Technologies

Réalisé par

ZARA VITA

Année universitaire : 2024/2025

SOMMAIRE

Remerciements

Tout d'abord, je rends grâce à **Allah, le Tout-Puissant et le Miséricordieux**, qui m'a accordé la santé, la patience et l'énergie nécessaires pour mener à bien ce parcours académique. Sans Sa guidance et Sa bénédiction, rien n'aurait été possible.

Je tiens à exprimer ma profonde gratitude à mes parents, **HAMISY ANDRIAMAHASOLO** et **MBOTIRIZIKY, dite Nozora**, qui ont toujours été mon premier soutien, tant moral que financier, depuis mes premiers pas à l'école maternelle jusqu'à l'aboutissement de ce Master. Leur amour, leurs sacrifices et leurs prières constantes sont la source de ma persévérance.

Mes remerciements vont également à ma famille résidant à Rabat, et en particulier à mon oncle **TOMBOHASY Ali BAKARY**, pour son accueil chaleureux, son aide précieuse et son soutien tout au long de mon séjour au Maroc.

Je souhaite remercier chaleureusement tous mes amis rencontrés durant mes années d'études au Maroc, qui ont su m'apporter motivation, solidarité et encouragements dans les moments difficiles comme dans les réussites.

J'adresse aussi mes sincères remerciements à mes professeurs de la **Faculté des Sciences et Techniques de Tanger**, dont l'enseignement et les conseils ont façonné ma formation académique. Je tiens à exprimer ma reconnaissance particulière au **Pr. Azmani Abdallah** pour m'avoir offert l'opportunité d'effectuer mon stage de fin d'études au sein de l'entreprise **SMART AUTOMATION TECHNOLOGIES**, ainsi qu'à ma tutrice de stage, **Dr. Ikhlass**, pour son encadrement rigoureux et bienveillant.

Enfin, mes remerciements les plus respectueux vont à **Pr. Bahij Meriem**, mon encadrante, pour son accompagnement constant, ses conseils précieux et sa disponibilité, sans lesquels ce travail n'aurait pas pu atteindre le niveau escompté.

À toutes celles et ceux qui, de près ou de loin, ont contribué à la réussite de ce travail, je dis du fond du cœur : **merci**.

Introduction Générale

La fidélisation des clients est aujourd'hui au cœur des stratégies de développement des entreprises opérant dans des environnements compétitifs. Dans de nombreux secteurs, l'acquisition d'un nouveau client se révèle significativement plus coûteuse que la rétention d'un client existant, tandis que la perte d'un client peut entraîner un impact direct et mesurable sur la rentabilité. Dès lors, la capacité à prédire le **churn** (attrition des clients) et à estimer la **Customer Lifetime Value (CLV)** constitue un enjeu stratégique majeur.

Avec l'essor des données massives et des outils de science des données, de nouvelles perspectives analytiques permettent d'anticiper non seulement quels clients sont susceptibles de partir, mais également **quand** ce départ risque de survenir et **quelle valeur économique** est en jeu. Cette vision tridimensionnelle (qui, quand, combien) est fondamentale pour orienter les décisions managériales et optimiser l'allocation des ressources.

Le présent mémoire s'inscrit dans cette dynamique en développant une approche intégrée de la gestion du churn, articulant plusieurs cadres méthodologiques complémentaires :

- **Le machine learning supervisé**, avec la mise en œuvre du modèle XGBoost, afin d'identifier les clients à risque à partir de leurs caractéristiques comportementales et transactionnelles.
- **L'analyse de survie**, mobilisant des outils non paramétriques (Kaplan–Meier), semi-paramétriques (Cox Proportional Hazards) et paramétriques (Weibull AFT), afin de prédire la temporalité du churn et de mettre en évidence les facteurs qui accélèrent ou ralentissent l'attrition.

- **Les modèles probabilistes BG/NBD et Gamma-Gamma**, permettant d'estimer la valeur vie client (CLV), en intégrant à la fois la fréquence d'achat future et la valeur monétaire moyenne des transactions.

L'originalité de ce travail réside dans la **combinaison cohérente de ces trois volets**, conduisant à la construction d'un dispositif décisionnel complet : détection précoce des clients à risque, anticipation du moment critique, et hiérarchisation selon la valeur économique.

Au-delà de l'apport scientifique, cette recherche se veut directement opérationnelle. Elle propose une segmentation des clients croisant risque de churn et valeur vie client, et débouche sur des recommandations stratégiques et tactiques pour guider les décisions de rétention. Enfin, des perspectives d'intégration dans des environnements de CRM et des systèmes de pilotage interactifs (dashboards, alertes automatisées) sont discutées, ouvrant la voie à une industrialisation durable de la gestion du churn.

Mots-clés : Churn client ; Fidélisation ; Analyse de survie : Weibull aft ; Machine Learning ; XGBoost ; BG/NBD ; Gamma-Gamma ; Customer Lifetime Value (CLV) ; Segmentation ; Data science appliquée.

Abstract

Customer churn prediction and management have become critical issues in competitive business environments, where retaining existing clients is significantly less costly than acquiring new ones. This study develops an integrated data-driven framework that combines machine learning, survival analysis, and probabilistic modeling to predict churn, understand its timing, and estimate the Customer Lifetime Value (CLV).

First, a supervised classification approach using **XGBoost** was implemented to accurately identify customers at risk of churn. Second, **survival analysis techniques** (Kaplan–Meier, Cox Proportional Hazards, Weibull AFT) were employed to predict the timing of churn events and to highlight the key factors influencing customer retention dynamics. Third, the **BG/NBD and Gamma-Gamma models** were applied to estimate the CLV, capturing both the expected purchase frequency and the average monetary value of transactions.

The joint use of these methods provides a comprehensive decision-making tool that answers three essential questions: *who* is likely to churn, *when* the churn may occur, and *how much* value is at stake. Empirical results show that approximately 39.5% of customers account for 80% of the overall CLV, underscoring the importance of targeted retention strategies.

Beyond methodological contributions, this research delivers actionable insights for businesses, including customer segmentation by risk and value, strategic retention recommendations, and the design of monitoring dashboards. While acknowledging the limitations related to data and model assumptions, the study demonstrates how advanced analytics can be transformed into measurable business gains through systematic validation and progressive industrialization.

Keywords: Customer churn; Retention; Survival analysis: Weibull AFT; Machine Learning; XGBoost; BG/NBD; Gamma-Gamma; Customer Lifetime Value (CLV); Segmentation; Applied Data Science, e-commerce.

Chapitre 1 : Contexte général du projet

INTRODUCTION

Présentation de l'organisme d'accueil

SMART AUTOMATION TECHNOLOGIES est une entreprise citoyenne, originellement fondée par un groupe de consultants seniors spécialistes des nouvelles technologies IT, du Data- Science et de l'Intelligence Artificielle. SMART AUTOMATION TECHNOLOGIES est résolument tournée vers l'innovation et la performance. Son équipe saura apporter toute son expertise et son savoir-faire aux profits des entreprises et des administrations qui souhaitent employer des solutions faisant réponse à leurs besoins et à leurs contraintes.

SMART AUTOMATION TECHNOLOGIES fait de l'innovation son crédo est investie en recherche et développement afin d'être à la pointe du progrès dans la technologie de l'Information et de l'intelligence artificielle. SMART AUTOMATION TECHNOLOGIES accompagne sereinement et efficacement ses clients en véritable partenaire dans l'élaboration, la mise en œuvre, la mise en place et l'évolution de leur système d'information et de leurs besoins en étude, en formation et en progiciels intégrés.

SMART AUTOMATION TECHNOLOGIES veut pérenniser sa relation avec ses clients en se plaçant comme un véritable partenaire et un maillon de leur chaîne de valeur, en leur apportant les compétences en informatiques, en apprentissage automatique nécessaires à leur bon fonctionnement et leur réussite an de leur permettre de se concentrer sur le cœur de leur métier.

Approches et valeurs

SMART AUTOMATION TECHNOLOGIES s'applique à cultiver une vision différente, elle donne une place importante à l'éthique et à la déontologie, ce qui lui garantit une relation de transparence et d'engagement avec ses clients.

SMART AUTOMATION TECHNOLOGIES entretient une relation de proximité fondée sur une confiance et une compréhension mutuelle pour mieux connaître les besoins et anticiper les attentes de ses clients.

SMART AUTOMATION TECHNOLOGIES est une Entreprise avec un bon potentiel humain, ses collaborateurs sont encadrés, suivis, évalués et formés de manière permanente.

SMART AUTOMATION TECHNOLOGIES possède les expertises des métiers de l'intégration, de l'environnement Infrastructure et de la Production en mode projet. La figure suivante représente les valeurs de l'entreprise :

Figure 1 : Valeurs de l'entreprise SAT

Métiers et services

SMART AUTOMATION TECHNOLOGIES, à travers ses différents métiers, (Services, Solutions dédiées, Support, Formations) met son expérience et son savoir-faire, dans la conception et la mise en œuvre des systèmes d'informations, au service de ses clients. SMART AUTOMATION TECHNOLOGIES prend en charge les projets de conception et d'accompagnement dans les différents secteurs d'activités où son expertise apporte une réelle valeur ajoutée.

La stratégie conseil de SMART AUTOMATION TECHNOLOGIES s'inscrit dans une démarche qui s'appuie sur une approche client, projet et méthode. Elle accompagne ses clients dans leurs projets d'organisation et d'évolution et conduit leur stratégie d'évolution en un plan d'orientation et de transformation de leur système d'information grâce à des opérations de :

- Gouvernance IT et Management de projets
- Conduite de Changement
- Élaboration d'un Schéma Directeur Informatique
- Urbanisation du système d'information
- Assistance à Maîtrise d'ouvrage
- Architecture logicielle
- Solution clé en main
- Benchmarking et audit logiciel

Produits et solutions

SMART AUTOMATION TECHNOLOGIES s'est attelée à la création de produits répondant à un besoin réel et s'est intéressée tout particulièrement à la création de solution tout en intégrant l'Intelligence artificielle, le data-Science et le Big data en particulier dans les domaines de l'industrie, de la logistique, du marketing et de la finance. SMART AUTOMATION TECHNOLOGIES s'intéresse également à la mise en place du processus de l'e-Transformation autour du principe d'écosystème digitale, intelligent et décisionnel (EcoDID).

SMART AUTOMATION TECHNOLOGIES développe actuellement plusieurs projets en recherche et développement en partenariat avec le milieu universitaire :

- Technologie Edge Box pour la maintenance prédictive
- Technologie Edge Data Center
- Plateforme logistique basée sur le concept EcoDID
- Plateforme marketing basée sur le concept EcoDID
- Plateforme éducative basée sur le concept EcoDID

Présentation du projet

Problématique

La gestion du churn client constitue aujourd'hui un défi majeur pour les entreprises, en raison de ses répercussions directes sur les revenus et de l'augmentation corrélative des coûts d'acquisition de nouveaux clients. De nombreuses études soulignent que conquérir un nouveau client est sensiblement plus onéreux que de conserver un client existant [1]. Ce constat a conduit au développement de stratégies spécifiques de prédiction et de rétention de l'attrition client [2], avec un accent particulier sur la fidélisation des clients présentant un fort retour sur investissement [3].

Au-delà de l'impact financier immédiat, la fidélisation s'avère également bénéfique sur le plan marketing, notamment par l'effet du bouche-à-oreille positif généré par des clients satisfaits [4]. La perte d'un client peut également engendrer une dynamique négative au sein de son cercle social, amplifiant ainsi le phénomène d'attrition [5]. Selon Gartner, le coût d'acquisition d'un nouveau client peut être jusqu'à cinq fois supérieur à celui de la fidélisation d'un client existant (Marketing Metrics), ce qui place la problématique de la rétention au cœur des stratégies de pérennisation des entreprises. De plus, selon Frederick Reichheld, une augmentation de seulement 5 % du taux de rétention peut entraîner une hausse des profits comprise entre 25 % et 95 % [6].

L'acquisition de nouveaux clients repose principalement sur des campagnes publicitaires coûteuses, des promotions attractives ou des incitations spécifiques. Bien qu'efficace pour soutenir la croissance à court terme, cette stratégie mobilise des investissements importants qui peuvent retarder l'atteinte de la rentabilité. À l'inverse, la fidélisation offre un modèle économique plus durable : les clients fidèles tendent à accroître leur fréquence d'achat, à générer

une valeur vie client (Customer Lifetime Value - CLV) plus élevée, et à promouvoir la marque de manière organique.

En ce sens, l'optimisation de la fidélisation permet d'initier un cercle vertueux : amélioration de la satisfaction, augmentation de la rétention, accroissement de la valeur client et renforcement de l'image de marque. Gartner estime par ailleurs que 80 % du chiffre d'affaires futur d'une entreprise proviendra de 20 % de ses clients existants, ce qui rend la rétention d'autant plus stratégique.

Cependant, la littérature et les pratiques industrielles révèlent plusieurs **limites dans les approches traditionnelles de prédiction du churn**. Les modèles de classification supervisée (comme la régression logistique, les arbres de décision ou XGBoost) permettent certes d'identifier les clients à risque, mais ils s'arrêtent à une vision binaire : « churn » ou « non churn ». Ils ne renseignent ni sur **le moment où le churn est susceptible de se produire**, ni sur **la valeur économique associée à la perte d'un client**. Or, dans un contexte décisionnel, ces dimensions temporelles et financières sont essentielles : savoir *quand* un client risque de partir et *combien il rapporte réellement* oriente directement la priorisation des actions marketing et budgétaires.

L'absence d'**analyse de survie** conduit à négliger le facteur temps, indispensable pour planifier des actions de rétention au moment opportun. De même, ignorer l'**estimation du Customer Lifetime Value (CLV)** limite fortement la capacité à hiérarchiser les interventions : toutes les pertes de clients ne se valent pas en termes d'impact financier.

Ces constats amènent à reformuler la problématique centrale de ce travail de recherche : comment construire une approche intégrée qui combine la **classification (qui prédit qui va partir)**, l'**analyse de survie (qui anticipe quand le départ est probable)** et l'**estimation du CLV (qui mesure combien vaut réellement chaque client)**, afin d'optimiser la rétention et la création de valeur ?

Ainsi, les questions fondamentales qui guident ce mémoire sont :

- Quels sont les facteurs les plus discriminants permettant d'identifier les clients à risque de churn (via des modèles de classification robustes comme XGBoost) ?
- Comment modéliser le facteur temps grâce aux approches de survie (Kaplan-Meier, CoxPH, Weibull AFT) pour anticiper le moment du départ et mieux calibrer les interventions ?
- Comment estimer la valeur vie client (CLV) grâce aux modèles BG/NBD et Gamma-Gamma afin de distinguer les clients stratégiques (VIP) des clients à faible valeur ?
- Comment combiner ces trois dimensions (qui, quand, combien) dans une **cartographie unifiée churn × CLV** afin de proposer des stratégies de rétention différenciées et économiquement rationnelles ?
- Quelles limites et perspectives se dégagent de cette approche intégrée (notamment en termes de qualité des données et de généralisabilité des modèles) ?

Ces interrogations structurent la problématique du présent mémoire, qui vise à dépasser les approches fragmentées pour offrir une vision holistique de la gestion du churn, conciliant **prédiction, temporalité et valeur économique**.

OBJECTIF DU PROJET

Objectif général du projet

L'objectif principal de ce travail est de proposer une approche intégrée, robuste et opérationnelle de la prédiction et de la gestion du churn, combinant **algorithmes de classification**, **analyse de survie** et **estimation de la valeur vie client (Customer Lifetime Value – CLV)**. Cette démarche vise à fournir aux décideurs une vision complète non seulement du risque de départ des clients, mais également du **moment probable de leur attrition** et de leur **valeur économique attendue**, afin d'optimiser les stratégies de fidélisation et de maximiser la rentabilité.

Objectif spécifique du projet

Pour atteindre cet objectif général, ce projet poursuit les finalités spécifiques suivantes :

1. **Détecter les clients à risque de churn** en mettant en œuvre un modèle de classification performant (XGBoost), afin de fournir un outil fiable de scoring opérationnel.
2. **Analyser la dimension temporelle du churn** à travers des modèles de survie (Kaplan–Meier, Cox Proportionnels, Weibull AFT), afin d'identifier les variables influençant la vitesse de départ des clients et d'anticiper le moment critique où des actions de rétention doivent être engagées.
3. **Estimer la valeur vie client (CLV)** en combinant les modèles probabilistes BG/NBD et Gamma-Gamma, permettant de segmenter les clients selon leur valeur économique attendue et de prioriser les efforts sur les clients « haute valeur / haut risque ».
4. **Élaborer une cartographie croisée Churn × CLV**, afin de proposer des stratégies différenciées de rétention (par exemple : forte valeur / faible risque, faible valeur / fort risque, etc.), garantissant une allocation optimale des ressources marketing et commerciales.
5. **Mettre en place un cadre méthodologique reproductible** incluant la validation croisée, le suivi par KPIs et la possibilité d'industrialisation (dashboards, intégration CRM), afin d'assurer la pérennité et l'opérationnalité des résultats obtenus.

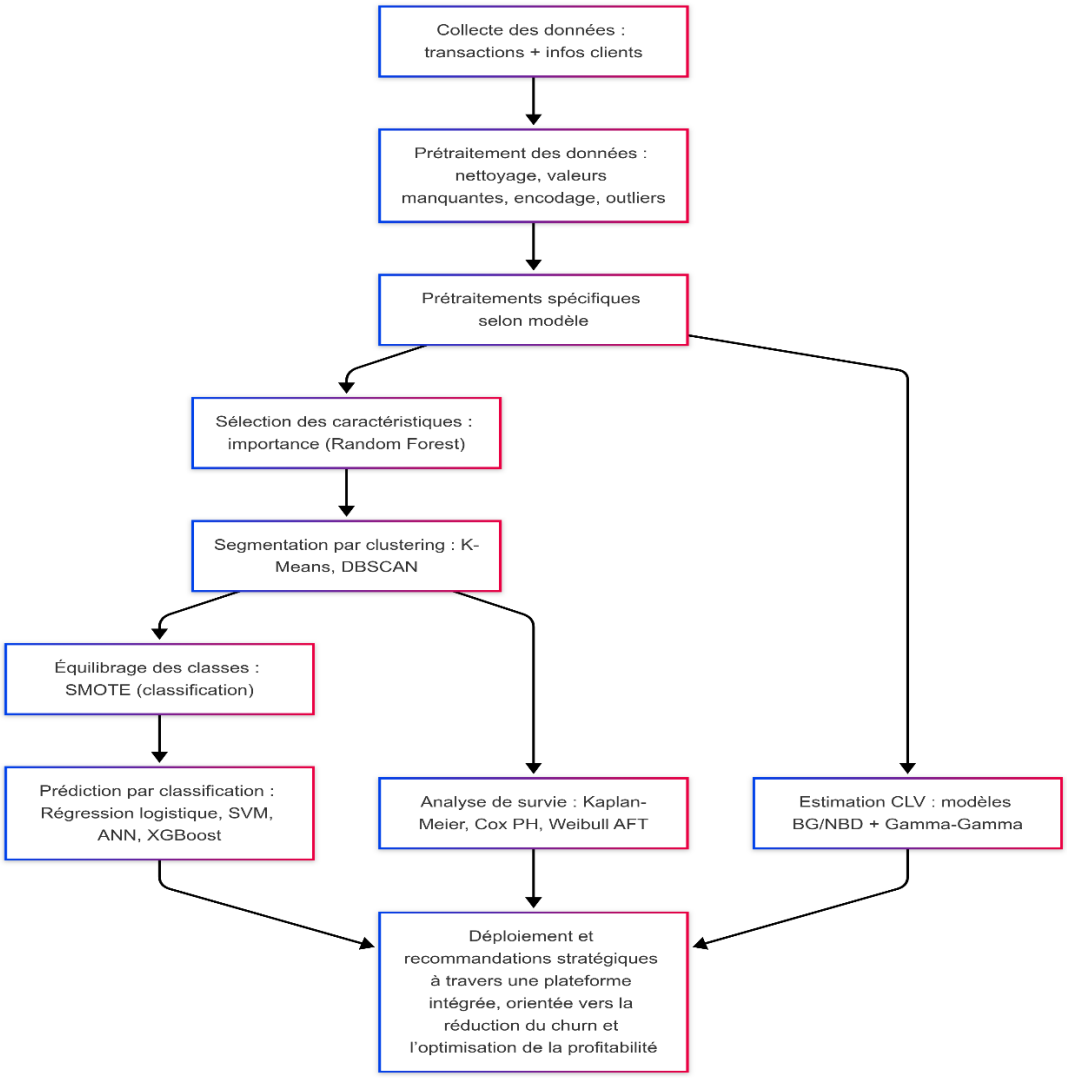
Approches méthodologiques

La méthodologie adoptée dans ce projet repose sur une succession d'étapes structurées, allant de la préparation des données jusqu'au déploiement final des modèles, afin d'offrir une vision complète et intégrée du phénomène de churn.

1. **Collecte des données**
   Les données sont issues de transactions et d'informations clients, comprenant des variables démographiques, comportementales et temporelles. Elles constituent la base d'analyse pour l'ensemble des modèles développés.

2. **Prétraitement des données**
   Une analyse exploratoire est effectuée afin d'identifier les tendances générales et la distribution des variables. Cette phase inclut le nettoyage des données, le traitement des valeurs manquantes, l'encodage des variables catégorielles et la détection des valeurs aberrantes.
   Il convient de noter que chaque famille de modèles nécessite des prétraitements spécifiques :
   - **Modèles de classification** : normalisation et équilibrage des classes.
   - **Analyse de survie** : les variables temporelles telles que *Tenure* ou *OrderCount* ne doivent pas être standardisées afin de conserver leur signification.
   - **Estimation de la Customer Lifetime Value (CLV)** : utilisation directe des données transactionnelles.
3. **Sélection des caractéristiques**
   Une étape de réduction dimensionnelle est réalisée grâce à l'importance des variables mesurée par les **forêts aléatoires (Random Forest)**. Cette étape vise à éliminer les variables peu pertinentes et à conserver uniquement les facteurs déterminants du churn.
4. **Segmentation par clustering**
   Avant la modélisation, une segmentation non supervisée est réalisée à l'aide de méthodes telles que **K-Means** et **DBSCAN**, afin d'identifier des profils clients distincts. Cette stratification permet d'adapter les modèles aux comportements spécifiques de chaque groupe et d'améliorer la précision prédictive. Comme montré par [2 - ancien], la segmentation préalable contribue significativement à la performance des modèles de prédiction.
5. **Équilibrage des classes**
   Dans le cas des modèles de classification (régression logistique, SVM, ANN et XGBoost), un rééquilibrage des données est effectué sur l'échantillon d'entraînement au moyen de la technique **SMOTE** (Synthetic Minority Over-sampling Technique).
6. **Prédiction par les modèles de classification**
   Plusieurs algorithmes sont implémentés et comparés, notamment la régression logistique, les SVM, les réseaux de neurones artificiels et XGBoost. Leurs performances sont évaluées selon des métriques classiques telles que la précision, le F1-score, l'exactitude et l'AUC.
7. **Analyse de survie**
   Pour capturer la dimension temporelle du churn, trois approches complémentaires sont appliquées :
   - Estimateur de Kaplan-Meier (KME),
   - Modèle semi-paramétrique de Cox (CoxPH),
   - Modèle paramétrique de Weibull AFT.
     Ces modèles permettent d'estimer la probabilité de départ d'un client dans un horizon donné (par exemple, à 3 mois).
8. **Estimation de la Customer Lifetime Value (CLV)**
   L'évaluation de la valeur vie client est réalisée à l'aide des modèles **BG/NBD** (Beta-Geometric/Negative Binomial Distribution) et **Gamma-Gamma**, permettant de quantifier la valeur financière future de chaque client.
9. **Déploiement et recommandations stratégiques**
   Les résultats issus des différentes approches (classification, survie et estimation de CLV) sont intégrés dans une plateforme unifiée, afin de fournir aux décideurs des recommandations opérationnelles et stratégiques exploitables en contexte réel.

La méthodologie suivie dans ce projet s'articule autour d'un ensemble d'étapes successives, allant de la préparation des données jusqu'à l'intégration finale des résultats dans des recommandations stratégiques. Le graphe ci-dessous illustre ce pipeline analytique, en montrant comment chaque composante (prétraitement, sélection des variables, segmentation, modélisation et estimation de la valeur vie client) contribue à l'objectif global de prédiction et de compréhension du churn.



Finalité du projet

La finalité de ce projet est de fournir à l'entreprise une **vision intégrée et quantitative de la fidélité et de la valeur de ses clients**, en combinant des méthodes mathématiques et d'intelligence artificielle. Plus précisément, le projet vise à :

- **Identifier les clients à risque de churn** de manière fiable, afin de cibler les actions de rétention et d'optimiser l'allocation des ressources marketing ;

- **Comprendre les déterminants et le timing du départ des clients** grâce à l'analyse de survie, permettant d'anticiper les comportements d'attrition et de mettre en place des interventions proactives ;
- **Estimer la valeur vie client (CLV) de manière probabiliste**, en intégrant la fréquence et le montant des transactions, pour prioriser les clients à forte contribution au chiffre d'affaires et guider les décisions stratégiques ;
- **Mettre en place des outils décisionnels exploitables**, tels que des segmentations, des cartes churn × CLV et des indicateurs de suivi, facilitant l'industrialisation des actions de fidélisation et l'optimisation continue des stratégies client.

Ainsi, ce projet ne se limite pas à la prédiction statistique du churn, mais propose **une approche globale combinant anticipation, quantification de la valeur et priorisation opérationnelle**, afin de soutenir durablement la performance commerciale et la rentabilité de l'entreprise.

Chapitre 2 : Cadrage logique du projet

2.1 Introduction

La dynamique actuelle de l'économie numérique et l'essor des plateformes d'e-commerce ont profondément transformé les comportements d'achat des consommateurs. Dans un contexte marqué par une concurrence accrue et une offre abondante, les clients deviennent de plus en plus volatils et moins enclins à maintenir une relation durable avec une marque donnée. Cette volatilité se traduit par le **phénomène de churn (attrition)**, qui constitue aujourd'hui l'un des principaux défis stratégiques pour les entreprises.

En effet, la perte de clients a un double impact : elle réduit directement le chiffre d'affaires et accroît les coûts liés à l'acquisition de nouveaux consommateurs. Or, de nombreuses études démontrent que fidéliser un client existant est beaucoup moins coûteux que d'en conquérir un nouveau, et qu'une amélioration même marginale du taux de rétention peut se traduire par une croissance substantielle de la rentabilité.

Ce constat rend indispensable la mise en place de dispositifs analytiques permettant non seulement de **prédire** quels clients risquent de partir, mais aussi de **comprendre** les facteurs qui expliquent ce départ et d'**estimer leur valeur économique future** afin de guider les efforts de rétention. Autrement dit, la problématique du churn ne se réduit pas à une question prédictive, mais s'inscrit dans une démarche plus large de **pilotage décisionnel et stratégique**.

C'est dans ce cadre que s'inscrit ce projet, qui mobilise conjointement des **outils mathématiques** et des **méthodes d'intelligence artificielle** afin de proposer une approche intégrée : segmentation des clients, classification des individus à risque, analyse de survie pour estimer le moment du churn, et modélisation probabiliste de la Customer Lifetime Value (CLV). Cette combinaison méthodologique vise à dépasser les limites des approches traditionnelles, souvent centrées uniquement sur la classification, pour offrir une vision plus riche et exploitable de la gestion de l'attrition client.

2.2 Définition des concepts clés

Avant de préciser la démarche méthodologique adoptée, il est nécessaire de clarifier certains concepts fondamentaux qui structurent le présent projet. Ces notions constituent le socle conceptuel à partir duquel s'articule le cadrage logique.

2.2.1 churn ou attrition client

Le churn, ou attrition client, désigne l'interruption de la relation commerciale entre un client et une entreprise. C'est généralement la **perte d'un client actif** sur une période donnée. Sa définition varie sensiblement selon les secteurs :

- Dans les industries **contractuelles** (banque, assurance, télécoms), le churn est explicite : un contrat est résilié, un service est interrompu.
- Dans les services SaaS, il est lié au non-renouvellement d'un abonnement.
- Dans les environnements **non-contractuels**, comme l'e-commerce, **le churn est implicite**, souvent identifié par une **inactivité transactionnelle** sur une période déterminée ou une absence de visites sur le site.

Le taux de churn constitue un indicateur clé de performance (*Key Performance Indicator – KPI*) pour les entreprises, car il traduit directement la capacité d'une organisation à fidéliser sa clientèle et à maintenir une base stable d'utilisateurs.

Typologies de churn

La compréhension fine des différentes formes de churn est essentielle pour élaborer des stratégies de rétention efficaces, tant. On distingue principalement deux grandes catégories de churn :

- **Churn volontaire** : résulte d'une décision délibérée du client de mettre fin à la relation commerciale. Les motifs incluent l'insatisfaction, l'attraction d'une offre concurrente plus avantageuse ou un changement de besoins.Batvoice
- **Churn involontaire** : survient indépendamment de la volonté du client, souvent en raison de circonstances externes telles que des problèmes de paiement (carte bancaire expirée, fonds insuffisants), des erreurs techniques ou des événements imprévus comme un déménagement ou un décès.

En complément, deux autres dimensions permettent de caractériser le churn :

- **Churn contractuel** : se manifeste par la résiliation explicite d'un contrat ou la non-reconduction d'un abonnement à son échéance. Ce type est fréquent dans les secteurs où les relations sont formalisées par des engagements contractuels, notamment en B2B .Stripe
- **Churn comportemental** : est inféré à partir d'une modification ou d'un arrêt du comportement du client, sans résiliation formelle. Il est souvent détecté par une inactivité prolongée, comme l'absence d'achats ou de visites sur une plateforme e-commerce.

Dans le secteur du e-commerce, caractérisé par des transactions ponctuelles et l'absence de contrats formels, le churn est majoritairement **comportemental** et **volontaire**. Les clients peuvent cesser leurs achats sans préavis ni justification, rendant la détection du churn plus complexe. Ce type de churn est influencé par des facteurs tels que l'expérience utilisateur, la satisfaction client et la concurrence. Il constitue donc un levier stratégique pour la fidélisation, justifiant l'adoption d'approches prédictives basées sur l'analyse comportementale.

Focus de l'étude

Cette étude se concentre principalement sur le **churn comportemental volontaire**. Ce choix est motivé par la possibilité d'anticiper ce type de churn à travers l'analyse des données transactionnelles et comportementales, permettant ainsi la mise en place de stratégies de rétention personnalisées et proactives. En identifiant les signaux faibles de désengagement, les entreprises peuvent intervenir avant la rupture de la relation client, optimisant ainsi la fidélisation et la rentabilité.

2.2.2 La satisfaction client

La satisfaction client se définit comme l’écart entre les attentes initiales d’un consommateur et la perception réelle de la qualité du service ou du produit reçu. Elle joue un rôle déterminant dans le processus de rétention : un client satisfait est plus susceptible de renouveler ses achats, de recommander la marque et de présenter une plus grande tolérance face à des incidents ponctuels.
Dans la littérature en marketing, la satisfaction est souvent considérée comme un **prédicteur direct de la fidélité**, mais également comme un facteur médiateur entre la qualité perçue et l’intention de ré-achat.

2.2.3 La fidélisation

La fidélisation dépasse la simple absence de churn. Elle renvoie à un engagement durable du client envers une marque, engagement qui se traduit à la fois par des comportements répétés (achats récurrents, recours régulier au service) et par une dimension affective (préférence, attachement, recommandation active).
La fidélisation est particulièrement stratégique, car un portefeuille de clients fidèles génère généralement une **valeur vie client (CLV)** plus élevée et une meilleure stabilité du revenu.

2.2.4 La Customer Lifetime Value (CLV)

La *Customer Lifetime Value* ou valeur vie client correspond à la valeur actualisée nette des profits générés par un client sur toute la durée de sa relation avec l'entreprise. Elle intègre à la fois :

- la **fréquence et le volume des achats**,
- la **durée estimée de la relation**,
- et les **marges associées**.

La CLV constitue un indicateur avancé, car elle permet d'orienter les stratégies de rétention en hiérarchisant les clients : tous les clients à risque de churn n'ont pas la même valeur stratégique pour l'entreprise. Ainsi, la CLV permet de distinguer les clients à forte valeur, qu'il convient de retenir prioritairement, de ceux dont la rentabilité est plus limitée.

2.2.5 L'analyse prédictive et l'intelligence artificielle

L'**analyse prédictive** regroupe l'ensemble des techniques statistiques et algorithmiques visant à anticiper des comportements futurs à partir de données historiques. Dans le cadre du churn, elle consiste à identifier, au moyen de modèles de classification ou de survie, les clients susceptibles de quitter l'entreprise dans un horizon temporel donné.
L'**intelligence artificielle (IA)**, et plus particulièrement le *machine learning*, occupe une place centrale dans ces approches, en permettant de traiter des volumes massifs de données hétérogènes et de détecter des patterns complexes souvent invisibles aux méthodes statistiques classiques.

2.2 Problématique et Questions de Recherche

Problématique principale

Dans un contexte concurrentiel où la fidélisation devient plus rentable que l'acquisition, une interrogation stratégique s'impose :
Comment identifier précocement les clients e-commerce à risque de churn volontaire, et

quelles actions personnalisées peuvent être déployées pour optimiser leur rétention de manière efficace et économiquement profitable ?

L'enjeu n'est donc pas uniquement prédictif, mais **décisionnel et financier** : il s'agit de cibler les clients dont la conservation génère un réel **retour sur investissement**, à l'aide de modèles non seulement précis, mais **optimisés pour maximiser la profitabilité**.

Objectif général

Ce mémoire vise à **concevoir une approche intégrée combinant modélisation mathématique et intelligence artificielle**, permettant de comprendre, anticiper et réduire le churn volontaire dans le cadre du e-commerce B2C. L'objectif central est d'identifier, à partir des données disponibles, les **déterminants comportementaux, transactionnels et contextuels** du départ des clients, puis de construire des modèles prédictifs **robustes, interprétables et directement exploitables** pour l'action opérationnelle.

L'ambition est de fournir un cadre méthodologique capable de formuler des **stratégies de rétention personnalisées**, alignées avec les objectifs de rentabilité et les impératifs de fidélisation à long terme. Cette démarche repose sur l'exploitation combinée de **données structurées et non structurées**, l'application d'**algorithmes supervisés**, de **méthodes de segmentation**, d'**analyse de survie** et de modèles probabilistes pour la **Customer Lifetime Value (CLV)**, ainsi que sur la mise en place d'**outils de pilotage dynamiques** permettant un suivi proactif du risque d'attrition.

(Pour le détail des objectifs spécifiques poursuivis dans le cadre de cette étude, le lecteur pourra se référer au Chapitre 1 – section "Objectifs spécifiques du projet".)

Questions de recherche

- Comment segmenter les clients e-commerce en fonction de leur **risque de churn et de leur valeur** afin d'identifier les segments stratégiques pour l'entreprise ?
- Quels modèles de **classification supervisée** permettent de prédire efficacement le churn, et quels sont les facteurs les plus discriminants pour chaque segment ?
- Comment l'**analyse de survie** (Kaplan-Meier, CoxPH, Weibull AFT) peut-elle être utilisée pour prédire **quand** les clients risquent de churner et identifier les variables qui accélèrent ou ralentissent ce départ ?
- Comment estimer la **Customer Lifetime Value (CLV)** à partir des modèles probabilistes (BG/NBD et Gamma-Gamma) et l'intégrer aux décisions de rétention pour prioriser les clients les plus rentables ?
- Comment combiner les résultats de segmentation, classification, analyse de survie et CLV pour formuler des **stratégies de rétention personnalisées**, optimisées en termes de valeur et de timing ?

2.3 Hypothèses de recherche.

L'hypothèse centrale de ce mémoire postule que :

« Le churn transactionnel volontaire dans le e-commerce B2C peut être anticipé et quantifié à partir de données transactionnelles, comportementales et contextuelles, et qu'une approche analytique intégrée – combinant segmentation, classification supervisée, analyse de survie et

estimation probabiliste de la valeur client (CLV) – permet de concevoir des stratégies de rétention ciblées, interprétables et économiquement rentables. »

Cette hypothèse principale se décline en plusieurs sous-hypothèses opérationnelles :

- **H1 – Facteurs transactionnels et comportementaux :** La fréquence, la récence, le montant moyen et la diversité des achats sont des prédicteurs significatifs du churn, et leur combinaison permet d'identifier de manière fiable les clients à risque élevé.
- **H2 – Apport de la segmentation :** La segmentation fine des clients basée sur les caractéristiques comportementales et transactionnelles (ex. clustering K-means) permet de détecter des profils homogènes de risque et de valeur, facilitant la personnalisation des actions de rétention.
- **H3 – Performance des modèles de classification :** L'utilisation d'algorithmes de type ensemble (XGBoost) permet d'obtenir une prédiction précise du churn tout en restant interprétable via des techniques explicatives (ex. SHAP), permettant de relier les résultats aux facteurs sous-jacents et de guider la prise de décision.
- **H4 – Analyse temporelle du churn :** Les méthodes d'analyse de survie (Kaplan–Meier, CoxPH, Weibull AFT) fournissent des estimations fiables du moment probable du churn et identifient les variables accélératrices ou retardatrices de l'attrition, offrant un calendrier optimal pour les interventions.
- **H5 – Estimation probabiliste de la valeur client :** L'application des modèles BG/NBD et Gamma-Gamma permet de quantifier la valeur future des clients, d'identifier les segments à forte CLV et d'allouer efficacement les ressources de rétention pour maximiser le retour sur investissement.
- **H6 – Intégration opérationnelle et optimisation économique :** La combinaison des modèles de segmentation, classification, analyse de survie et estimation CLV fournit un cadre robuste pour la **décision stratégique et opérationnelle**, permettant de concilier performance prédictive, optimisation des ressources marketing et rentabilité financière.
- **H7 – Validation et robustesse :** La fiabilité des prédictions et des recommandations dépend de la qualité des données et de la méthodologie employée, incluant la validation croisée, la calibration des modèles et la surveillance continue des performances dans le temps.

2.4 Arbre à Problèmes

La mise en place d'une stratégie de rétention efficace exige en amont une identification précise des problèmes qui affectent la relation client. L'outil de l'**arbre à problèmes** est particulièrement pertinent, car il permet de représenter de manière structurée la logique causale reliant les **causes profondes** aux **conséquences observées**, en passant par le **problème central**.

2.4.1 Problème central

Le problème au cœur de ce projet est formulé comme suit :
« Le churn volontaire et la baisse de satisfaction réduisent la rentabilité et compromettent la pérennité des entreprises de e-commerce. »

Ce problème résume la difficulté majeure rencontrée par les acteurs du secteur : malgré l'acquisition régulière de nouveaux clients, une part significative d'entre eux interrompt prématurément sa relation, entraînant un manque à gagner substantiel.

2.4.2 Causes principales

Les causes qui alimentent ce problème central peuvent être regroupées en plusieurs dimensions :

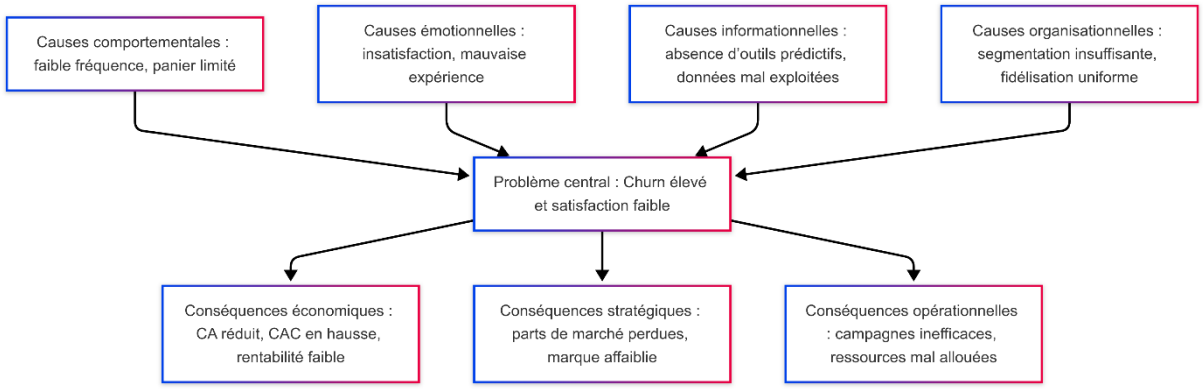
1. **Causes comportementales** : faible fréquence d'achat, baisse de récence, panier moyen limité, diversification des achats insuffisante.
2. **Causes émotionnelles** : insatisfaction liée à la qualité du service, mauvaise expérience client, absence de personnalisation, retards de livraison.
3. **Causes informationnelles et analytiques** : absence d'outils prédictifs performants, incapacité à exploiter les données transactionnelles et comportementales de manière optimale, faible intégration des données non structurées (avis, feedbacks).
4. **Causes organisationnelles** : manque de segmentation fine, stratégies de fidélisation uniformes, absence d'alignement entre les priorités marketing et la valeur réelle des clients (CLV).

2.4.3 Conséquences

Les effets négatifs liés au churn et à la baisse de satisfaction sont multiples et interdépendants :

- **Conséquences économiques** : diminution du chiffre d'affaires récurrent, hausse des coûts d'acquisition (CAC) pour compenser les pertes, rentabilité réduite.
- **Conséquences stratégiques** : perte de parts de marché au profit des concurrents, affaiblissement du capital-marque, fragilisation de la croissance à long terme.
- **Conséquences opérationnelles** : sur-sollicitation des équipes marketing sur des campagnes de réacquisition peu efficaces, allocation inefficiente des ressources de rétention.

2.4.4 Schéma de l'arbre à problèmes



2.5 Arbre à Objectifs

L’**arbre à objectifs** constitue la transposition positive de l’arbre à problèmes. Il permet de reformuler les difficultés identifiées en objectifs à atteindre, en mettant en évidence les liens de causalité entre les moyens (racines) et les fins (branches). Dans le cadre de ce projet, il s’agit de passer d’un constat de churn élevé et de satisfaction faible vers un ensemble de solutions visant à renforcer la rétention, optimiser la rentabilité et soutenir la croissance durable.

2.5.1 Objectif central

L’objectif principal du projet peut être formulé ainsi :
« Réduire le churn volontaire et améliorer la satisfaction client grâce à des modèles analytiques intelligents et interprétables, afin d’optimiser la rétention et la profitabilité dans le e-commerce. »

Cet objectif constitue le pivot autour duquel s’articulent les autres sous-objectifs.

2.5.2 Objectifs intermédiaires

Les causes du problème central (cf. 2.3) sont reformulées en objectifs intermédiaires, regroupés selon les mêmes dimensions :

1. **Objectifs comportementaux**

    - Encourager la fréquence et la récence des achats.
    - Stimuler l’augmentation du panier moyen et la diversification des produits consommés.

2. **Objectifs émotionnels**

    - Améliorer l’expérience client et la personnalisation des interactions.
    - Réduire l’insatisfaction liée aux retards, litiges ou défauts de service.

3. **Objectifs informationnels et analytiques**

    - Développer des modèles prédictifs robustes (segmentation, classification, analyse de survie).
    - Exploiter les données structurées et estimer la **Customer Lifetime Value (CLV)** pour prioriser les clients stratégiques.
    - Intégrer les données non structurées (avis, feedbacks) afin d’enrichir les signaux de churn et affiner la détection précoce.

4. **Objectifs organisationnels**

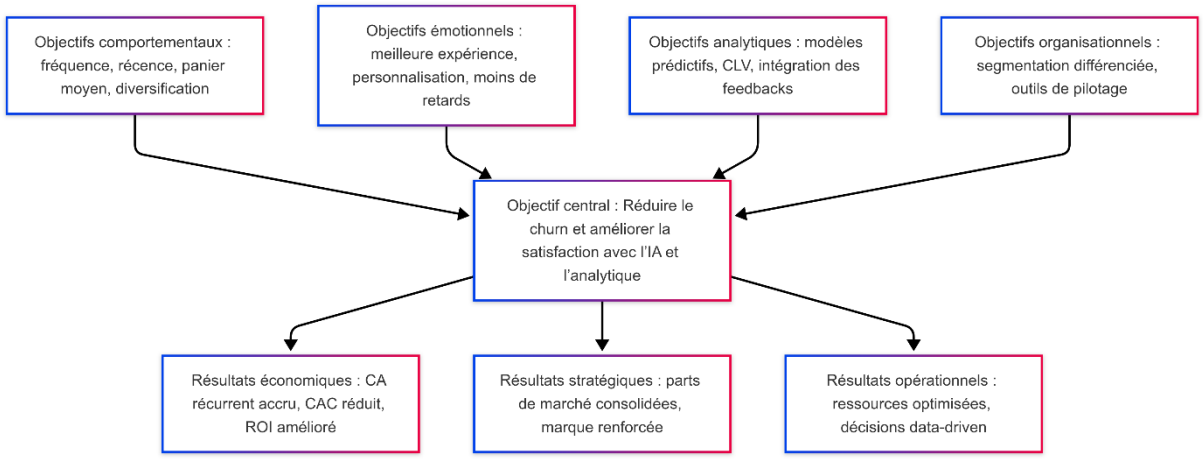
    - Mettre en place une segmentation différenciée permettant des actions de rétention ciblées.
    - Développer des outils de pilotage et de suivi permettant une gestion proactive du risque d’attrition.

2.5.3 Résultats attendus

La mise en œuvre de ces objectifs devrait conduire à plusieurs bénéfices :

- **Résultats économiques** : amélioration du chiffre d'affaires récurrent, réduction du coût d'acquisition, maximisation du retour sur investissement des campagnes de fidélisation.
- **Résultats stratégiques** : consolidation des parts de marché, renforcement du capital-marque, meilleure compétitivité à long terme.
- **Résultats opérationnels** : allocation optimisée des ressources marketing, adoption d'une approche fondée sur la donnée (data-driven decision making).

2.4.4 Schéma de l'arbre à objectifs



2.6 Arbre à Solutions

L'**arbre de solutions** constitue la traduction opérationnelle de l'arbre à objectifs. Il met en évidence les moyens et approches méthodologiques mobilisables pour atteindre les objectifs définis précédemment. Chaque solution proposée correspond à une réponse concrète à un objectif spécifique, et l'ensemble s'inscrit dans une logique intégrée permettant d'aboutir à l'objectif central de réduction du churn et d'amélioration de la satisfaction client.

2.6.1 Solution centrale

La solution globale du projet peut être formulée comme suit :
« Développer une approche analytique hybride combinant modèles mathématiques, techniques d'intelligence artificielle et outils de suivi décisionnel, afin de prédire, comprendre et gérer le churn client tout en estimant la valeur vie client (CLV). »

2.6.2 Solutions intermédiaires

1. **Solutions comportementales**

- Déployer des campagnes de rétention personnalisées basées sur la segmentation des clients (ex. offres ciblées pour les clients à forte CLV).
- Mettre en œuvre des programmes de fidélité et de stimulation (récompenses, promotions adaptées).

2. **Solutions émotionnelles**

- Intégrer l'analyse des feedbacks clients pour améliorer l'expérience utilisateur.
- Développer des mécanismes proactifs de résolution des litiges et retards afin de limiter l'insatisfaction.

3. **Solutions analytiques et méthodologiques**

- Appliquer des modèles de **segmentation et classification** afin d'identifier les clients à risque de churn.
- Utiliser des techniques d'**analyse de survie** (Kaplan-Meier, Cox PH, Weibull AFT) pour estimer la durée de vie client et anticiper les moments critiques de désengagement.
- Mettre en place des modèles probabilistes de type **BG/NBD** pour prédire la fréquence d'achats futurs.
- Coupler ces approches avec le modèle **Gamma-Gamma** pour estimer la profitabilité individuelle et calculer la **Customer Lifetime Value (CLV)**.

4. **Solutions organisationnelles et stratégiques**

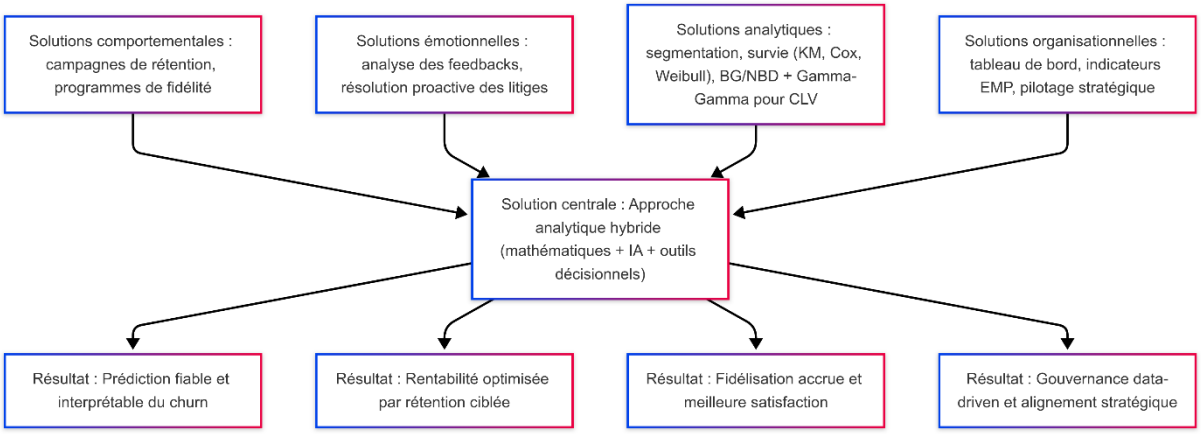
- Développer un tableau de bord interactif (Power BI ou équivalent) pour le suivi du churn, de la satisfaction et de la CLV en temps réel.
- Intégrer des indicateurs orientés profit (Expected Maximum Profit – EMP) dans l'évaluation des modèles, afin d'aligner la prédiction avec les priorités économiques.

2.6.3 Résultats attendus des solutions

La mise en place des solutions ci-dessus devrait permettre :

- Une **prédiction fiable et interprétable** du churn, facilitant la prise de décision.
- Une **optimisation de la rentabilité** via des actions de rétention ciblées sur les segments les plus stratégiques.
- Une **fidélisation accrue** grâce à une meilleure compréhension des comportements et des attentes des clients.
- Une **gouvernance data-driven**, où les décisions marketing et stratégiques s'appuient sur des modèles rigoureux et des indicateurs de performance pertinents.

2.6.4 Schéma de l'arbre de solutions



2.7 Cadre Conceptuel

Le cadre conceptuel de ce projet vise à structurer et articuler les concepts clés, les variables étudiées et les relations attendues entre elles, afin de guider l'analyse empirique et l'application des modèles prédictifs et probabilistes.

Concepts clés

- **Churn client (attrition)** : Dans le contexte e-commerce, le churn désigne le départ volontaire d'un client, c'est-à-dire l'interruption de ses achats ou interactions avec l'entreprise sur une période donnée. La compréhension et la prédiction du churn permettent de cibler les clients à risque et de maximiser le retour sur investissement des actions de fidélisation.
- **Segmentation client** : La segmentation consiste à regrouper les clients en sous-populations homogènes selon leur comportement d'achat, leur valeur économique et leur risque de churn. Elle constitue une étape préalable essentielle pour adapter les stratégies de rétention.
- **Analyse de survie** : L'analyse de survie permet d'estimer la probabilité qu'un client continue à rester actif jusqu'à un instant donné, et de déterminer les facteurs accélérant ou retardant le churn. Elle apporte une dimension temporelle indispensable pour planifier les interventions marketing.
- **Modèles prédictifs et apprentissage automatique** : Les algorithmes de classification supervisée (comme XGBoost) sont utilisés pour estimer la probabilité de churn de chaque client à partir de variables transactionnelles et comportementales. L'interprétabilité des modèles, via les méthodes SHAP ou équivalentes, est intégrée afin de rendre les décisions exploitables.
- **Customer Lifetime Value (CLV)** : La CLV représente la valeur économique attendue d'un client sur un horizon donné. L'estimation probabiliste (BG/NBD + Gamma-Gamma) permet de prioriser les clients selon leur potentiel financier futur, et de calibrer les investissements marketing de manière optimale.

Variables et relations attendues

Le modèle conceptuel du projet repose sur l'idée que le churn est influencé par plusieurs catégories de variables :

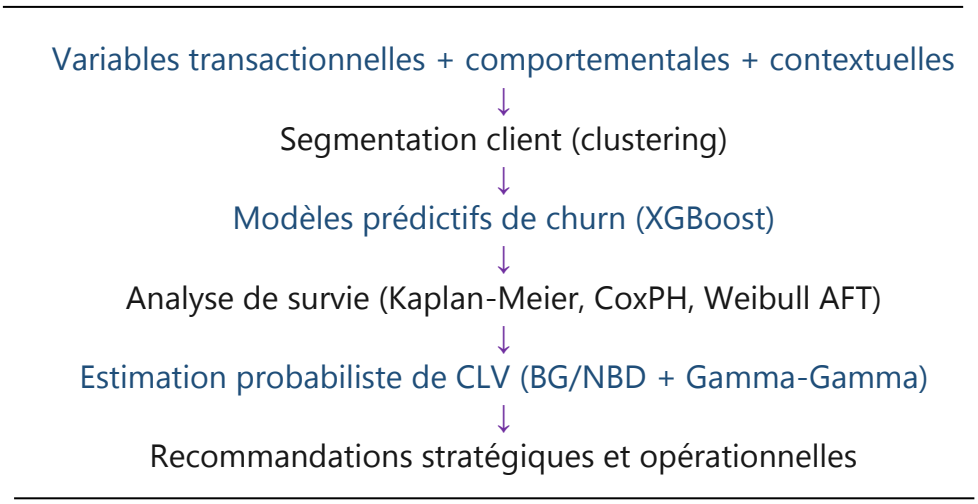
1. **Variables transactionnelles** : fréquence d'achat, récence, montant moyen, diversité des produits achetés.
2. **Variables comportementales et contextuelles** : interactions avec le service client, réponses aux promotions, avis et feedbacks.
3. **Variables de valeur** : CLV estimée, panier moyen cumulé, potentiel de contribution au chiffre d'affaires futur.

Ces variables sont supposées interagir de manière à prédire :

- La probabilité de churn d'un client à un instant donné (modèles de classification).
- Le moment attendu de churn (analyse de survie).
- La valeur future attendue du client (modèles probabilistes CLV).

Schéma conceptuel

Un schéma conceptuel synthétise la logique suivante :



2.8 Périmètre et Limites

Le présent travail se concentre sur l'analyse et la prédiction du churn transactionnel volontaire dans un contexte e-commerce B2C, avec une extension possible aux clients professionnels dans certains cas. L'étude repose sur l'exploitation de données historiques transactionnelles et comportementales, incluant la récence, la fréquence et le montant des achats, ainsi que des

indicateurs contextuels tels que les interactions clients et, lorsque disponible, des informations issues de la relation client.

Les principaux axes couverts par ce projet comprennent :

- **Segmentation des clients** pour identifier des groupes homogènes en termes de risque de churn et de valeur potentielle, via des techniques de clustering (ex. K-means).
- **Classification prédictive** pour déterminer la probabilité de churn de chaque client, en s'appuyant sur des algorithmes supervisés (XGBoost) et des méthodes d'interprétabilité (SHAP).
- **Analyse de survie** pour estimer le moment probable du churn et identifier les variables accélératrices ou retardatrices.
- **Estimation probabiliste de la valeur vie client (CLV)** à l'aide des modèles BG/NBD et Gamma-Gamma, afin de prioriser les actions de rétention sur les clients à forte valeur économique.
- **Recommandations stratégiques et opérationnelles**, basées sur l'intégration des résultats des différentes analyses pour soutenir la prise de décision et optimiser le retour sur investissement des actions de fidélisation.

Limites du projet

Malgré la rigueur méthodologique, certaines limites doivent être explicitement reconnues :

1. **Limites liées aux données**
    - Les données disponibles sont principalement transactionnelles et comportementales, avec un accès limité aux données qualitatives ou textuelles (avis clients, interactions sur réseaux sociaux).
    - Certaines informations importantes pour le churn, telles que la satisfaction exacte ou les motivations subjectives des clients, sont approximées par des proxies (ex. notes, plaintes, nombre de retours).
    - Les données historiques peuvent présenter des biais de sélection ou d'agrégation, limitant la généralisation des résultats à de nouvelles populations ou périodes.
2. **Limites liées aux modèles**
    - Les modèles de classification et de survie reposent sur des hypothèses statistiques et probabilistes (indépendance des événements, distributions paramétriques), qui peuvent être violées dans certains sous-groupes de clients.
    - L'estimation de la CLV via BG/NBD et Gamma-Gamma suppose une stabilité du comportement d'achat dans le futur, ce qui peut ne pas refléter des changements soudains du marché ou des stratégies concurrentielles.
3. **Limites opérationnelles**
    - L'application des recommandations dépend de la capacité de l'entreprise à collecter, intégrer et exploiter les données dans un système CRM ou un environnement décisionnel opérationnel.
    - La mise en œuvre d'actions personnalisées nécessite des ressources marketing et financières, ce qui peut restreindre la portée pratique des stratégies proposées.
4. **Limites temporelles et contextuelles**

- Les prédictions sont valables sur un horizon défini et peuvent nécessiter des recalibrages périodiques pour rester pertinentes face aux évolutions du comportement client ou des conditions du marché.
- Le projet se concentre sur le e-commerce B2C et, dans une moindre mesure, B2B, et ne couvre pas d'autres secteurs où la dynamique du churn peut différer significativement.

En résumé, ce projet offre une approche intégrée, robuste et méthodologiquement rigoureuse pour comprendre, anticiper et gérer le churn client, tout en reconnaissant les limites liées aux données, aux modèles et au contexte opérationnel. Ces limites fournissent également des pistes pour des recherches futures et des améliorations méthodologiques.

Précaution épistémologique

L'étude adopte une posture d'**exploration et de démonstration appliquée**, visant à illustrer le **potentiel des approches analytiques avancées** dans le cadre d'un projet de lutte contre le churn client. Elle **n'a pas vocation à prédire le churn avec une précision absolue**, mais à proposer un **cadre reproductible, modulaire et intelligible**, susceptible d'être répliqué dans un contexte réel avec des données internes spécifiques.

Conclusion

En dépit de ses limites, cette étude offre une **base solide pour toute entreprise souhaitant initier ou renforcer une stratégie de rétention data-driven**, en combinant rigueur scientifique, faisabilité technique et pertinence stratégique. Elle pose également les fondements d'éventuels **travaux de recherche appliquée ou d'implémentation professionnelle** en environnement réel, notamment autour de la fidélisation personnalisée et de la modélisation comportementale.

Chapitre 3 : Revue de littérature et veille scientifique sur la prédiction du churn et la valeur vie client

3.1 Introduction

La prédiction et la gestion de l'attrition client (*churn*) constituent aujourd'hui un champ de recherche central, à la croisée de la statistique appliquée, de l'intelligence artificielle et du marketing relationnel. La multiplication des données transactionnelles et comportementales, combinée à l'intensification de la concurrence, a conduit la communauté scientifique à développer des modèles toujours plus sophistiqués pour comprendre, anticiper et réduire ce phénomène.

Dans cette perspective, la littérature académique s'est structurée autour de trois axes majeurs. Le premier concerne les **modèles de classification supervisée**, qui visent à identifier les clients à risque de départ à partir de leurs caractéristiques historiques. Le second relève de l'**analyse de survie**, empruntée initialement à la biométrie et à la fiabilité, et qui permet de répondre à une question souvent négligée par les approches purement prédictives : *quand* le client est-il susceptible de quitter l'entreprise. Enfin, un troisième axe s'est développé autour de l'**estimation de la Customer Lifetime Value (CLV)**, indicateur financier clé permettant de hiérarchiser les efforts de rétention en fonction de la valeur économique future des clients.

Parallèlement, la **veille technologique** met en évidence l'émergence d'outils et de bibliothèques de plus en plus accessibles (Python, R, plateformes BI, environnements cloud) qui favorisent l'industrialisation de ces modèles, tout en ouvrant la voie à des approches intégrées mêlant classification, survie et CLV.

Ce chapitre se propose donc de présenter un état de l'art structuré sur la prédiction du churn et la modélisation de la valeur vie client. Il mettra en lumière les principales contributions théoriques et méthodologiques, tout en soulignant les limites et lacunes persistantes, afin de situer la contribution spécifique de ce mémoire.

3.2 Revue de littérature sur le churn

3.2.1 Définitions et typologies du churn

Le **churn** (ou attrition) désigne le phénomène par lequel un client cesse sa relation commerciale avec une entreprise. Il peut prendre plusieurs formes, en fonction du type de marché et de la nature du contrat liant le client à l'entreprise. On distingue généralement :

- le **churn volontaire**, lorsqu'un client choisit de se désabonner ou de ne plus utiliser un service ;
- le **churn involontaire**, lorsqu'il découle de facteurs externes ou techniques (ex. : faillite, déménagement, décès, problèmes de paiement) ;
- le **churn contractuel**, qui concerne les environnements où l'abonnement est formellement défini (ex. télécommunications, banque, assurances) ;
- le **churn non contractuel**, où l'inactivité ou la baisse d'interactions tient lieu d'attrition implicite (ex. e-commerce, streaming).

Cette typologie est essentielle car elle détermine les méthodes de détection et de prédiction adaptées à chaque secteur.

3.2.2 Enjeux stratégiques du churn dans le e-commerce et les services

Avec l'adoption généralisée des systèmes de gestion de la relation client (**CRM**) tels que Salesforce, HubSpot et Oracle, les entreprises disposent désormais d'outils performants leur permettant de mesurer, analyser et anticiper leur taux de désabonnement. La problématique du churn est devenue un enjeu central, particulièrement dans les industries fonctionnant sur un modèle d'abonnement (banque, assurance, télécommunications, e-commerce).

La littérature met en évidence le **coût considérable du churn** : acquérir un nouveau client coûte jusqu'à **10 fois plus cher** que de conserver un client existant [3-2-2-article 2]. De plus, la valeur économique d'une entreprise est directement liée au nombre de clients actifs, ce qui influence sa rentabilité, sa trésorerie et sa capacité d'investissement. Des études ont montré que les clients fidèles sont plus rentables à long terme, d'où l'intérêt de stratégies de **Customer Lifetime Value (CLV/CLTV)** visant à mesurer et optimiser cette rentabilité [3-2-4-article 2].

Ainsi, l'analyse du churn s'inscrit non seulement dans une logique de **préservation du portefeuille client**, mais également dans une dynamique de **création de valeur durable** pour l'entreprise.

3.2.3 Travaux académiques sur le churn : secteurs et approches

La recherche académique sur le churn est fortement marquée par les grandes entreprises opérant à grande échelle, notamment dans les télécommunications [3-2-1]–[3-2-3], le commerce de détail en ligne [3-2-4] et le secteur bancaire [3-2-5]. Ces industries, qui desservent des millions de clients, offrent un terrain riche pour le développement et l'évaluation de modèles prédictifs.

En revanche, les **petites entreprises** restent sous-représentées dans la recherche sur l'attrition, alors même qu'elles sont confrontées à des défis spécifiques : ressources limitées, bases de

données plus petites, dépendance accrue à chaque client. Cette asymétrie souligne la nécessité de développer des approches adaptées aux **PME** et aux environnements à faible volumétrie de données.

La littérature identifie plusieurs causes majeures du churn :

- facteurs **économiques** (prix, promotions concurrentielles),
- facteurs **relationnels** (qualité du service, satisfaction, expérience client),
- facteurs **technologiques** (pannes, obsolescence, transition vers des services digitaux).

Pour y répondre, des modèles de prédiction basés sur des méthodes statistiques et d'apprentissage automatique ont été largement explorés : régressions logistiques, arbres de décision, forêts aléatoires, réseaux de neurones et, plus récemment, deep learning.

3.2.4 Gestion de la Relation Client (CRM) et modèles d'attrition

La **Gestion de la Relation Client (CRM)** regroupe l'ensemble des processus et systèmes mis en place pour établir des relations durables et rentables avec les clients, soutenant ainsi les stratégies commerciales [3-2-7-article 3]. Dans les télécommunications comme dans le e-commerce, la fidélisation constitue l'une des activités clés du CRM.

L'objectif principal est d'instaurer une relation **gagnant-gagnant** entre le client et l'entreprise, en s'appuyant sur quatre dimensions :

1. **Identification du client**
2. **Attraction du client**
3. **Fidélisation du client**
4. **Développement du client**

Ces dimensions reposent sur des techniques de **data mining** variées :

- associations,
- classification,
- clustering,
- prévision,
- régression,
- découverte de séquences.

Dans ce cadre, les **modèles d'attrition** constituent des techniques de classification permettant de prédire les départs clients et d'orienter les politiques de fidélisation [3-2-8-article 3].

3.3 Approches classiques de prédiction du churn (classification)

3.3.1 Régression logistique

La régression logistique est l'une des méthodes les plus utilisées historiquement pour la prédiction du churn. Elle permet de modéliser la probabilité qu'un client quitte l'entreprise en fonction de variables explicatives comme la fréquence d'achat, la récence, ou la durée de la relation client. Son principal avantage réside dans sa **robustesse et son interprétabilité**, ce qui en a fait un modèle de référence dans les premiers travaux sur l'attrition. Toutefois, cette méthode suppose une relation linéaire entre les variables explicatives et le log-odds de la variable cible, ce qui peut constituer une limite dès lors que les données révèlent des relations complexes et non linéaires. Cette faiblesse a progressivement conduit les chercheurs et praticiens à se tourner vers des méthodes plus flexibles.

3.3.2 Méthodes d'apprentissage automatique (Machine Learning)

Avec l'explosion des volumes de données et la diversité croissante des comportements clients, les approches classiques se sont révélées insuffisantes. Une revue de littérature [Ahmad et al., 2019] recense **plus de 200 études menées entre 2013 et 2023**, mettant en évidence une évolution marquée des modèles statistiques traditionnels vers des méthodes de machine learning, hybrides et profondes.

• Arbres de décision

Les arbres de décision figurent parmi les techniques prédictives les plus répandues, en raison de leur simplicité d'interprétation, de leur capacité à générer des règles explicites et de leur robustesse dans des contextes commerciaux [168]. Leur principe repose sur une division récursive de l'espace des données en sous-groupes homogènes selon des critères de sélection comme le gain d'information, l'entropie ou l'impureté de Gini.

Les variantes les plus connues incluent **C4.5** et **C5.0**, largement utilisées pour segmenter la clientèle en groupes à risque ou fidèles [170]. L'algorithme **CART (Classification and Regression Tree)** constitue une autre approche, basée sur la minimisation de l'impureté entre les nœuds parents et enfants. Dans la prédiction du churn, plusieurs études empiriques ont validé leur efficacité : par exemple, Al-Najjar et al. [143] ont montré que **C5.0 surpassait d'autres modèles** comme les réseaux de neurones, les arbres bayésiens ou CHAID sur des données financières, tandis qu'Alizadeh et al. [200] ont mis en évidence la supériorité de **CTree** en termes de précision, F-mesure et AUC.

Malgré ces résultats positifs, les arbres de décision montrent des limites face à la complexité des relations non linéaires. Cela a favorisé le recours à des méthodes d'ensemble, visant à améliorer la performance prédictive [170], [171].

• Méthodes d'ensemble

Les méthodes d'ensemble combinent plusieurs modèles de base pour produire une prédiction plus robuste et plus précise. Elles peuvent être homogènes (plusieurs arbres de décision, par exemple) ou hétérogènes (mélangeant arbres, SVM, ANN, etc.).

Dans la littérature, les plus utilisées sont **Random Forest**, **Gradient Boosting Trees (GBT)**, **XGBoost**, **AdaBoost**, **CatBoost** et **LightGBM** [91], [118], [120]. Ces approches se sont révélées particulièrement performantes dans les télécommunications, la finance et le jeu mobile. Parmi elles, **Random Forest** est souvent cité pour sa robustesse face au bruit et au surapprentissage [105], et surpasse dans de nombreux cas la régression logistique, les SVM et

même les réseaux de neurones [76], [77]. Le **boosting** (notamment XGBoost et LightGBM) a également montré d'excellents résultats, parfois supérieurs à Random Forest [125].

Des approches hybrides ont également été proposées, combinant plusieurs méthodes d'ensemble (par exemple Random Forest + AdaBoost) ou associant différents types de modèles, afin d'accroître la précision globale [113]. Par ailleurs, [35] propose un modèle hybride **logit leaf**, combinant arbres de décision et régression logistique, qui offre un compromis entre performance et interprétabilité.

Néanmoins, la principale critique adressée à ces méthodes reste leur **manque d'interprétabilité**, ce qui a conduit à l'intégration d'outils explicatifs tels que **SHAP** ou **LIME** [125].

• Réseaux de neurones et deep learning

Depuis 2015, les réseaux de neurones et l'**apprentissage profond** se sont imposés comme une approche prometteuse pour prédire le churn, surpassant régulièrement les méthodes traditionnelles [167]. Les architectures comme les **ANN, CNN, RNN, LSTM** ou encore les **Transformers** permettent de capturer des dynamiques complexes, notamment temporelles, et d'obtenir des gains de performance allant jusqu'à **15 %** par rapport aux modèles classiques.

Certaines études ont proposé des architectures hybrides (par exemple CNN + LSTM bidirectionnel), ou encore des prétraitements par **clustering** (PPFCM) avant classification, améliorant considérablement les performances. D'autres travaux ont exploré des méthodes bio-inspirées (chasse au cerf, chauves-souris, troupeau d'éléphants) pour optimiser les hyperparamètres. Enfin, des approches innovantes ont consisté à transformer des séries temporelles en images pour les classifier avec des CNN pré-entraînés.

Malgré ces avancées, le principal défi reste la **faible explicabilité** des réseaux de neurones. Pour y répondre, des techniques explicatives comme **LIME**, **SHAP** ou les mécanismes d'**attention** ont été intégrées afin de rendre les prédictions plus compréhensibles [167].

3.3.2 Synthèse sur les approches classiques et justification du choix des modèles

Les approches classiques de modélisation du churn se sont largement appuyées sur des méthodes statistiques et d'apprentissage automatique dites « traditionnelles ». Parmi elles, la **régression logistique** constitue l'algorithme de référence, offrant une interprétation claire des coefficients et une bonne robustesse sur des données tabulaires. D'autres méthodes, telles que les **machines à vecteurs de support (SVM)** et les **réseaux de neurones artificiels (ANN)**, ont été explorées dans la littérature, chacune apportant des avantages spécifiques en termes de performance prédictive ou de capacité à capturer des relations non linéaires.

Dans ce travail, bien que le **modèle XGBoost** ait initialement été retenu comme choix principal, l'expérience menée a consisté à implémenter et comparer plusieurs algorithmes à savoir régression logistique, SVM, ANN et XGBoost afin d'évaluer leurs performances respectives. Les résultats seront comparés selon des métriques standard de classification en contexte de churn, à savoir : **précision (Precision)**, **F1-Score**, **exactitude (Accuracy)** et **aire sous la courbe ROC (AUC)**. Cette démarche permet d'assurer une évaluation rigoureuse et objective des performances des différents modèles, tout en confirmant la supériorité attendue de XGBoost.

Justification du choix de XGBoost

Le modèle **XGBoost** a été retenu en raison de sa supériorité prédictive avérée, avec un gain moyen de **+8,2 % d'AUC** sur les benchmarks churn par rapport aux modèles classiques comme la régression logistique ou le SVM [1-4]. Plusieurs caractéristiques renforcent ce choix:

- **Efficacité computationnelle** : entraînement de 3 à 10 fois plus rapide et inférence jusqu'à 100× plus rapide que les ANN sur données tabulaires [2-4].
- **Explicabilité intégrée** : grâce aux valeurs **SHAP**, XGBoost fournit une interprétation des facteurs influents du churn, contrairement aux ANN qui se comportent comme des boîtes noires [3-4].
- **Gestion native des données hétérogènes et manquantes**, réduisant considérablement les besoins de prétraitement [4-4].
- **Faible consommation mémoire**, ce qui facilite son déploiement en temps réel et en environnement de production [5-4].
- **Robustesse face aux déséquilibres de classes**, grâce à une régularisation adaptée et à une pondération efficace, deux aspects essentiels dans le contexte du churn [1-4].

Ainsi, bien que plusieurs modèles aient été testés et comparés, **XGBoost constitue le choix final privilégié**, combinant à la fois performance prédictive, efficacité computationnelle et facilité d'interprétation, ce qui en fait une solution particulièrement adaptée au problème étudié.

3.4 Analyse de survie appliquée au churn

Les approches classiques de prédiction du churn, telles que la régression logistique ou les arbres de décision, se sont révélées efficaces pour classifier les clients en « churners » ou « non-churners ». Toutefois, elles présentent une limite essentielle : elles ne permettent pas d'estimer **le moment de l'attrition**. Or, dans de nombreux secteurs (télécoms, finance, e-commerce, paiement en ligne), la question n'est pas uniquement de savoir qui va se désabonner, mais également **quand** cet événement est susceptible de survenir. C'est précisément ce défi que l'analyse de survie vise à relever, en modélisant le **temps jusqu'à un événement** et en tenant compte de phénomènes de **censure** et de **troncature** souvent présents dans les données [3-4-3]. Comme l'indique les articles [3-4-X1] et [3-4-X2], les modèles de survie sont naturellement adaptés à la modélisation du désabonnement client.

3.4.1. Origines et bases méthodologiques

Historiquement, l'analyse de survie a émergé dans les domaines médical et actuariel afin de modéliser la durée de survie des patients ou la fiabilité de systèmes techniques. Une première avancée fut la méthode non paramétrique de **Kaplan–Meier (1958)**, permettant d'estimer empiriquement une fonction de survie à partir de données censurées. Toutefois, cette approche reste descriptive et ne permet pas d'intégrer directement des covariables explicatives.

Un tournant majeur a été l'introduction du **modèle à risques proportionnels de Cox (1972)**, qui repose sur une approche semi-paramétrique. Ce modèle combine la souplesse d'une

estimation non paramétrique de la fonction de risque de base et la possibilité d'évaluer l'effet de covariables via des coefficients interprétables, tout en tenant compte de la censure [3-4-3]. Ce caractère hybride explique sa longévité et son adoption massive dans des domaines variés, allant de la biomédecine au marketing.

En parallèle, les **modèles paramétriques** ont connu un essor important. Regroupés sous l'appellation **Accelerated Failure Time (AFT)**, ils postulent que les covariables agissent directement sur le temps jusqu'à l'événement, par un effet multiplicatif. Ces modèles incluent plusieurs distributions de durées :

- **Exponentielle**, adaptée à un risque constant mais souvent trop restrictive ;
- **Weibull**, très flexible, capable de modéliser des risques croissants ou décroissants ;
- **Log-normale**, utile lorsque les temps de survie suivent une distribution asymétrique ;
- **Log-logistique**, permettant de capturer des risques qui augmentent puis diminuent dans le temps ;
- **Gompertz**, fréquemment utilisée pour des phénomènes de vieillissement ou d'attrition progressive.

Parmi ces distributions, la **loi de Weibull** occupe une place singulière. En effet, elle constitue le seul cas pouvant être formulé à la fois dans le cadre **AFT (Accelerated Failure Time)** et dans celui des **modèles à risques proportionnels (PH)**. Cette dualité conceptuelle en fait un véritable pont entre les deux grandes familles de modèles de survie et lui confère une **flexibilité pratique et théorique** rare [27-4 ; 28-4 ; 29-4]. Ainsi, le Weibull est couramment utilisé non seulement pour sa robustesse empirique, mais aussi parce qu'il facilite la comparaison et la transition entre les paradigmes AFT et PH.

3.4.2. Évolutions modernes : machine learning et deep learning

À partir des années 1990, plusieurs travaux ont cherché à dépasser les limites du modèle de Cox, notamment en intégrant des structures non linéaires via des réseaux de neurones (Faraggi & Simon, 1995 ; Xiang et al., 2000). Bien que ces premiers modèles n'aient pas surpassé Cox de manière significative, ils ont ouvert la voie à une intégration progressive de l'apprentissage automatique dans l'analyse de survie [3-4-4][3-4-26].

L'introduction des **Random Survival Forests (RSF)** par Ishwaran et al. (2008) a constitué une autre étape clé, permettant de modéliser des effets d'interaction complexes et non linéaires sans hypothèse stricte de risques proportionnels [3-4-8][3-4-15]. Plus récemment, les méthodes issues de l'apprentissage profond ont donné naissance à des variantes puissantes :

- **DeepSurv** (Katzman et al., 2018) : extension neuronale du modèle de Cox, performante mais encore contrainte par l'hypothèse des risques proportionnels [3-4-9].
- **Deephit** (Lee et al., 2018) : approche innovante basée sur une distribution discrète du temps d'événement, adaptée à la prédiction du moment exact du churn [3-4-17].
- **Cox-Time** (Kvamme et al., 2019) : version neuronale flexible du modèle de Cox, levant la contrainte des risques proportionnels et améliorant l'efficacité computationnelle [3-4-15].

- **LSTM appliqués à l'analyse de survie** : utilisés pour capturer des séquences transactionnelles et l'évolution dynamique du comportement client (Yiwen et al., 2021), notamment dans le secteur des paiements en ligne.

Ces méthodes offrent une meilleure capacité prédictive dans des contextes complexes et dynamiques, mais au prix d'un coût computationnel plus élevé et d'une moindre interprétabilité.

3.4.3. Applications sectorielles au churn

L'analyse de survie a été appliquée dans divers secteurs pour comprendre et prédire le churn :

- **Télécommunications** : Masarifoglu & Buyuklu (2019) utilisent des modèles de survie (Kaplan–Meier, Cox) pour identifier les facteurs influençant la durée de rétention des clients, incluant la tarification, la durée de contrat et les campagnes promotionnelles. Les résultats permettent d'optimiser les stratégies de rétention proactive [Masarifoglu & Buyuklu, 2019].
- **Banque et assurance** : plusieurs travaux appliquent des modèles de Cox et AFT pour analyser la durée de relation bancaire et l'attrition dans l'assurance, en mettant en évidence l'importance des comportements transactionnels et des facteurs démographiques [3-4-24].
- **Paiements en ligne / fintech** : une étude récente (Özalpay & Taşkın, 2024) applique l'analyse de survie à un grand acteur turc des paiements en ligne. Les auteurs montrent que les variables liées aux taux de remboursement, aux changements de commissions et aux volumes de transaction influencent fortement le risque de churn, avec un pic critique après 25 mois de relation. Le modèle caractérisant le churn comme une absence d'activité d'un mois s'est révélé le plus efficace [Özalpay & Taşkın, 2024].
- **E-commerce** : Yiwen et al. (2021) appliquent des réseaux LSTM combinés à des modèles de survie pour prédire l'attrition dans les paiements en ligne, confirmant que la dynamique transactionnelle (flux séquentiels de paiement, contexte marchand) est déterminante pour anticiper le moment exact du churn.

3.4.4. Limites et perspectives

Malgré leur pertinence, les méthodes de survie appliquées au churn rencontrent plusieurs limites dans le contexte e-commerce :

- les données sont **non contractuelles**, ce qui rend difficile la définition précise de la date d'attrition (souvent approximée par une absence d'activité) ;
- la granularité temporelle est parfois insuffisante pour estimer des fonctions de risque fiables ;
- les modèles avancés (DeepHit, Cox-Time, LSTM) exigent des **volumes massifs de données séquentielles**, rarement disponibles en pratique.

En dépit de ces contraintes, l'analyse de survie apporte une contribution essentielle en permettant de **quantifier le risque temporel d'attrition**, et en offrant des outils décisionnels précieux pour optimiser les stratégies de fidélisation dans des environnements hautement compétitifs comme les télécoms, la finance ou le paiement en ligne.

3.4.5 Synthèse

L’analyse de survie appliquée au churn repose sur une hiérarchie méthodologique allant des estimateurs descriptifs aux modèles semi-paramétriques et paramétriques. Dans le cadre de ce mémoire, trois approches complémentaires ont été considérées : l’Estimateur de Kaplan-Meier, la régression de Cox et les modèles de durée paramétriques de type Weibull AFT.

Premièrement, **l’Estimateur de Kaplan-Meier (KME)** a été utilisé afin d’obtenir la courbe de survie de la population dans son ensemble et de fournir une première estimation non paramétrique de la probabilité de rétention. Le KME est reconnu comme un outil de référence dans les études de survie, mais il demeure insuffisant dès lors que l’on souhaite analyser l’effet des caractéristiques individuelles (récence, fréquence, montant, etc.) sur le risque de churn. C’est pourquoi il a servi de point de départ, avant de passer à des modèles explicatifs.

Deuxièmement, la **régression de Cox** a été mobilisée comme méthode centrale de ce mémoire. Ce choix se justifie à plusieurs niveaux :

- Elle constitue une **référence historique et théorique** en analyse de survie, ayant prouvé sa robustesse dans de multiples contextes empiriques [3-4-15].
- Elle conserve un **avantage d’interprétabilité** majeur, car les coefficients estimés restent directement interprétables, contrairement aux réseaux neuronaux [3-4-26].

Malgré la montée des méthodes avancées (DeepHit, Cox-time, Gradient Boosting), plusieurs études récentes montrent que la régression de Cox conserve des performances prédictives **comparables à ces modèles plus sophistiqués**, notamment sur la métrique IBS (Integrated Brier Score) et le C-index [3-4-28].

- De plus, les travaux de Kevin Singh (2024) confirment que Cox dépasse DeepHit et Cox-Time en termes de C-index et IBS lorsque l’échantillon est limité (≈ 1 000 observations). Ce n’est qu’à partir de volumes beaucoup plus importants (10 000 à 15 000 observations) que ces modèles avancés deviennent supérieurs, ce qui met en évidence leur dépendance à de vastes ensembles de données – une contrainte non compatible avec les ressources disponibles dans ce mémoire.

Troisièmement, un **modèle paramétrique de type Weibull AFT** a également été testé. Ce modèle est fréquemment mobilisé dans la littérature de churn, car il repose sur une hypothèse explicite concernant la distribution du temps d’attrition, hypothèse qui peut être validée ou infirmée empiriquement. Par ailleurs, il a été montré que les résultats du modèle Weibull AFT pouvaient être similaires à ceux de la régression de Cox, notamment dans l’étude appliquée aux petites entreprises B2B réalisée par Hills Jr. et al. (University of Virginia, 2021). Cette complémentarité a permis de renforcer la robustesse des analyses en comparant un modèle semi-paramétrique et un modèle paramétrique.

En revanche, certaines approches alternatives ont été volontairement écartées. Les **forêts aléatoires de survie (RSF)**, introduites par Ishwaran et al. (2008) puis étendues au cadre de risques concurrents (Ishwaran et al., 2014), présentent un intérêt certain dans le domaine médical ou en présence de multiples événements compétitifs. Toutefois, dans le contexte étudié – un cadre transactionnel non contractuel et focalisé sur l’attrition volontaire – leur complexité computationnelle, leur moindre interprétabilité et l’absence de risques concurrents pertinents limitent leur apport opérationnel. Leur utilisation a donc été jugée non prioritaire.

En résumé, l'évolution des modèles de survie traduit un élargissement progressif des outils disponibles, allant de l'estimation purement descriptive (Kaplan–Meier) aux cadres semi-paramétriques (Cox) et paramétriques (AFT), avec le **modèle de Weibull** comme pivot méthodologique offrant une interprétation à la fois en termes de temps d'événement et de proportionnalité des risques.

Cette combinaison permet à la fois de bénéficier de la solidité des approches classiques et d'assurer une applicabilité pratique dans le cadre des données réelles disponibles.

3.5 Estimation du Customer Lifetime Value (CLV)

3.5.1. Panorama des familles de modèles

La prédiction de la valeur vie client (Customer Lifetime Value, CLV) a fait l'objet de recherches abondantes et a donné naissance à un ensemble diversifié de familles méthodologiques. Gupta et al. (2006) distinguent six grandes catégories :

- **les modèles RFM** (récence, fréquence, montant),
- **les modèles probabilistes** (famille Buy-Till-You-Die, BTYD),
- **les modèles économétriques** (par exemple chaînes de Markov),
- **les modèles de persistance**,
- **les modèles issus de l'intelligence artificielle**,
- **les modèles de diffusion et de croissance**.

Ces approches présentent des finalités et des contraintes très différentes : certaines sont essentiellement descriptives (RFM, diffusion), d'autres orientées vers la prédiction (probabilistes, économétriques), et d'autres encore visent une automatisation à grande échelle (IA).

Toutefois, dans le cadre du **commerce électronique non contractuel**, plusieurs études empiriques récentes (Jasek et al., 2018 ; 2019) convergent vers un constat : seules les approches probabilistes et, dans une moindre mesure, économétriques offrent des performances robustes pour un environnement caractérisé par :

- des achats **irréguliers**,
- des montants de panier **fortement variables**,
- et l'absence d'un contrat explicite avec le client.

Les méthodes RFM, trop statiques, ne permettent pas une projection fiable à long terme. Les modèles de persistance, souvent naïfs, reposent sur des extrapolations linéaires. Les approches d'intelligence artificielle (réseaux neuronaux, arbres de décision complexes, deep learning) montrent un fort pouvoir prédictif mais posent trois problèmes majeurs : besoin de données massives, difficultés d'interprétation (« black box ») et coûts computationnels élevés. Enfin, les modèles de diffusion sont conçus pour étudier l'adoption de nouveaux produits, et non la valeur générée par une base client existante [Gupta et al., 2006 ; Jasek et al., 2019].

Conséquence méthodologique pour ce mémoire : la revue se concentre sur les deux familles jugées les plus pertinentes : les modèles probabilistes (BTYD) et, à titre de comparaison, les modèles économétriques (chaînes de Markov).

3.5.2. Approches économétriques : chaînes de Markov et limites pratiques

Les modèles économétriques de type chaîne de Markov appliqués à la CLV décrivent le cycle de vie client comme une suite d'**états comportementaux**, et estiment les probabilités de transition entre ces états au cours du temps. L'intérêt de cette approche est double :

1. elle permet de **visualiser les trajectoires comportementales** (passages successifs entre états) ;
2. elle facilite l'**expérimentation de scénarios de rétention** en simulant l'effet d'actions marketing sur les transitions (Jasek et al., 2018).

Cependant, deux limites majeures réduisent leur applicabilité dans ce projet :

- elles requièrent des **variables contextuelles riches** (sources de trafic, localisation, canaux marketing) pour segmenter les états de manière pertinente ;
- elles exigent un **volume historique conséquent** et une granularité temporelle fine pour estimer une matrice de transition stable.

Dans notre cas, l'absence de variables contextuelles détaillées et la profondeur limitée de l'historique disponible ne permettent pas d'exploiter pleinement cette approche. En conséquence, les modèles de chaîne de Markov sont écartés, malgré leur intérêt théorique.

3.5.3. Famille probabiliste (BTYD) : principes généraux

Les modèles **Buy-Till-You-Die (BTYD)** constituent la référence pour l'estimation du CLV en contexte non contractuel. Leur principe est de distinguer :

- la **fréquence d'achat**, généralement modélisée par une loi de type binomiale négative (NBD),
- le **processus d'attrition implicite** (churn non contractuel), modélisé par une loi de durée de vie (Pareto, Beta-géométrique, etc.).

Ces modèles incorporent également l'**hétérogénéité individuelle** via des distributions a priori (Gamma, Beta), ce qui leur confère une souplesse adaptée à la diversité des comportements clients (Fader et al., 2005 ; Schmittlein et al., 1987).

L'information clé obtenue est la **probabilité qu'un client soit encore actif (« alive »)** à un instant donné, ainsi que l'espérance de ses achats futurs. Ces indicateurs sont centraux pour projeter le CLV individuel.

3.5.4. Principales variantes et critères de sélection

La littérature distingue deux grandes catégories de modèles probabilistes :

- **Modèles estimés par Maximum de Vraisemblance (MLE)** : plus rapides et adaptés aux grands volumes :
- **Modèles Bayésiens / MCMC** : plus flexibles mais coûteux computationnellement :

Jasek et al. (2019) ont réalisé une comparaison exhaustive de onze modèles probabilistes sur des datasets massifs de e-commerce (2.3 millions de clients), les classant en ces deux catégories, notamment :

- **Modèles de Maximum de Vraisemblance (MLE)** : Plus simples et plus rapides à estimer.
  - **NBD (Negative Binomial Distribution)** : Modèle de base supposant un taux d'achat constant et hétérogène, mais sans modélisation de la défection (**Ehrenberg, 1959**).
  - **Pareto/NBD** : Le modèle historique et de référence. Il modélise à la fois le taux de transaction (distribué Gamma) et le temps de vie actif (distribué Pareto). Bien que performant, son estimation est réputée complexe (**Schmittlein et al., 1987**).
  - **BG/NBD (Beta-Geometric/NBD)** : Introduit par **Fader, Hardie & Lee (2005)**, il simplifie le processus de défection du Pareto/NBD (en utilisant une distribution Beta-Géométrique) pour surmonter ses complexités computationnelles tout en conservant un pouvoir prédictif équivalent, voire supérieur dans certains cas.
  - **MBG/NBD (Modified BG/NBD)** : Une extension du BG/NBD proposée par **Batislam et al. (2007)**.
  - **BG/CNBD-k & MBG/CNBD-k** : Variantes introduisant une régularité fixe dans les achats (**Platzer, 2016**).

Modèles Bayésiens (MCMC) : Plus flexibles mais computationally intensifs, nécessitant une estimation par méthodes MCMC (Markov Chain Monte Carlo).

  - **Pareto/NBD (HB)** : Version Hiérarchique Bayésienne du Pareto/NBD (**Ma & Liu, 2007**).
  - **Pareto/NBD (Abe)** et **Pareto/NBD (Abe M2)** : Variantes d'**Abe (2009)** permettant d'incorporer des covariables.
  - **Pareto/GGG** : Modèle sophistiqué à trois distributions Gamma (**Platzer & Reutterer, 2016**).

Les modèles bayésiens (Pareto/NBD (HB), Pareto/NBD (Abe), Pareto/NBD (Abe M2), Pareto/GGG), bien que très flexibles, sont extrêmement lourds à calculer pour de grands volumes de données et requièrent une expertise avancée pour leur mise en œuvre et leur diagnostic. L'article de **Jasek et al. (2019)** note d'ailleurs que la variante d'Abe a sous-performé

sur de multiples critères. Leur complexité opérationnelle est disproportionnée pour un premier déploiement de modèle CLV dans notre contexte.
Le modèle **NBD** est écarté car il ne modélise pas la défection, ce qui est crucial dans un environnement non contractuel.
Les modèles **MBG/NBD** et les variantes à régularité fixe (**BG/CNBD-k**, **MBG/CNBD-k**) sont des extensions plus complexes du BG/NBD. Si elles apportent des améliorations théoriques, **Jasek et al. (2019)** concluent que le **BG/NBD standard offre une performance stable et excellente** sans ajouter cette complexité supplémentaire.

3.5.5. Complément monétaire : le modèle Gamma–Gamma

La dimension monétaire du CLV est souvent estimée via le modèle **Gamma–Gamma** (Fader & Hardie), qui repose sur l'hypothèse que la dépense moyenne par client suit une loi Gamma conditionnée à une moyenne inobservée. En pratique, le modèle permet d'estimer, pour chaque individu, la valeur monétaire moyenne future de ses transactions.

Combiné au BG/NBD (qui prédit la fréquence et la probabilité de survie), le Gamma–Gamma fournit une **estimation complète du CLV individuel**, intégrant à la fois le volume et la valeur des transactions.

3.5.6. Synthèse et justification méthodologique

Notre choix se porte sur le modèle **BG/NBD (Beta-Geometric Negative Binomial Distribution)**, dont la pertinence est largement confirmée par la littérature académique et les applications empiriques.

Tout d'abord, le BG/NBD est particulièrement adapté au **contexte du e-commerce**, où les relations entre entreprises et clients sont généralement **non contractuelles** et où la défection n'est pas directement observable. Contrairement à d'autres cadres théoriques, ce modèle capture efficacement cette incertitude inhérente à la dynamique transactionnelle.

En matière de **performance prédictive**, Jasek et al. (2019), dans une étude comparative menée sur de vastes ensembles de données e-commerce, concluent que « de bons résultats stables ont été obtenus avec les modèles BG/NBD et Pareto/NBD », tout en mettant en évidence des « améliorations significatives par rapport au modèle de référence Status Quo ». Ces résultats soulignent la robustesse et la fiabilité du BG/NBD dans des environnements variés.

Un autre atout réside dans sa **simplicité computationnelle**. Comme le rappellent Fader, Hardie et Lee (2005), le BG/NBD a été conçu afin de surmonter les complexités numériques du Pareto/NBD, tout en conservant un **niveau de précision équivalent**. Il s'impose ainsi comme un compromis particulièrement efficace entre rigueur théorique et facilité d'estimation.

De plus, ce modèle s'avère **parfaitement adapté aux données disponibles** dans ce projet. Il ne requiert que des données transactionnelles minimales à savoir identifiant client, date de transaction et montant sans recourir à des variables explicatives supplémentaires, souvent indisponibles ou coûteuses à collecter.

Concernant l'**implémentation pratique**, le BG/NBD bénéficie de solutions robustes et bien documentées dans l'écosystème Python, qui constitue aujourd'hui l'environnement de référence en data science appliquée. En particulier, la bibliothèque ***lifetimes*** représente une ressource mature et largement adoptée pour la mise en œuvre des modèles BTYD. Elle offre une API intuitive pour l'estimation, la prédiction et l'évaluation du modèle BG/NBD, tout en s'intégrant naturellement dans la pile logicielle standard (pandas, numpy, scikit-learn, Jupyter). Cette intégration facilite la reproductibilité des résultats, la validation expérimentale et le déploiement opérationnel. L'abstraction des détails mathématiques permet par ailleurs une utilisation en quelques lignes de code, rendant le modèle à la fois **puissant et accessible**.

Enfin, la **stabilité statistique** du BG/NBD est soulignée par Jasek et al. (2019), qui mettent en évidence la faiblesse des écarts-types associés à ses estimations. Cet élément renforce la confiance dans les prédictions et confirme la pertinence du modèle comme choix initial pour une modélisation fiable de la valeur vie client.

En résumé, le BG/NBD se distingue par sa capacité à combiner **précision prédictive, robustesse statistique, simplicité computationnelle et facilité d'implémentation**. Ces caractéristiques en font un choix optimal pour ce projet, garantissant à la fois une cohérence méthodologique et une applicabilité directe dans le cadre de la prédiction du CLV en e-commerce.

3.6 XVeille technologique

- Outils logiciels et bibliothèques utilisées dans la recherche académique et en entreprise :
  - Python (scikit-learn, lifelines, btgym, PyMC3).
  - R (survival, BTYD, caret).
  - Outils BI (Power BI, Tableau, Streamlit).
- Innovations récentes : AutoML, Deep Learning, modèles bayésiens appliqués au CLV.
- Cas concrets d'utilisation par des grandes entreprises (Amazon, Netflix, Spotify).

3.8 Synthèse et positionnement du projet

L'examen de la littérature révèle plusieurs limites structurelles dans les approches existantes du churn. Premièrement, la majorité des travaux se concentrent sur la **classification** des clients (par exemple via des modèles de régression logistique, SVM ou réseaux de neurones), sans intégrer d'autres dimensions analytiques [3-2-1]–[3-2-5]. Deuxièmement, les approches combinant simultanément **analyse de survie** et **estimation de la Customer Lifetime Value (CLV)** demeurent rares, alors même que ces deux perspectives sont essentielles pour une compréhension dynamique et financière du phénomène d'attrition. Enfin, on constate une **faible articulation entre performance algorithmique et impact économique**, les études privilégiant souvent la maximisation d'indicateurs techniques (précision, AUC, etc.) au détriment d'une analyse des retombées stratégiques pour l'entreprise.

Le présent projet vise à combler ces limites en adoptant une approche intégrée qui articule :

- une **segmentation préalable par K-Means**, afin de mettre en évidence des profils de clients distincts et d'adapter les modèles aux comportements hétérogènes ;
- une **phase de classification** reposant sur plusieurs modèles (régression logistique, SVM, ANN et XGBoost). Bien que XGBoost ait été identifié comme choix final en raison de ses performances supérieures [1-4], l'évaluation comparative reste nécessaire afin de sélectionner le modèle le plus robuste dans le contexte étudié ;
- une **analyse de survie** mobilisant le modèle de Cox Proportionnal Hazards (CoxPH) et le modèle paramétrique Weibull AFT. Ces deux modèles seront comparés à travers des métriques établies telles que le *C-index*, le *Brier Score* et l'*Integrated Brier Score (IBS)* ;
- une **estimation du CLV** reposant sur le modèle BG/NBD, retenu pour son équilibre entre puissance prédictive et interprétabilité dans les environnements non contractuels [Fader et al., 2005 ; Jasek et al., 2020].

L'apport de ce travail réside dans sa **double orientation scientifique et opérationnelle**. Sur le plan académique, il s'agit d'un cadre méthodologique complet combinant segmentation, classification, survie et CLV, rarement implémenté conjointement dans les études existantes. Sur le plan managérial, cette démarche répond directement aux besoins des entreprises en intégrant la prédiction de l'attrition à une évaluation de l'impact financier, ce qui reste très peu documenté dans la pratique professionnelle [Verbraken et al., 2014].

Ainsi, ce mémoire propose une **démarche unifiée et pragmatique**, capable de produire des résultats à la fois rigoureux et exploitables par les décideurs. Pour une présentation détaillée de l'approche méthodologique globale et de son enchaînement, se référer au chapitre 1 (*Approches méthodologiques*).

Chapitre 4 : Fondements théoriques et techniques analytiques des modèles.

4.1 Algorithmes de classification

4.1.1 Théorie de l'algorithme XGBoost

L'**Extreme Gradient Boosting** (XGBoost) est une amélioration du **Gradient Tree Boosting** introduit par Friedman. Il repose sur le principe de combinaison de plusieurs apprenants faibles (arbres de décision) pour former un classifieur robuste, selon une approche additive et itérative. Ce qui distingue XGBoost des versions classiques réside dans ses **optimisations internes**, notamment l'intégration de techniques de **régularisation explicite**, permettant de mieux contrôler la complexité du modèle et de limiter le surapprentissage.

Soit un ensemble d'exemples d'apprentissage $\{(x_1,y_1),\ldots,(x_n,y_n)\}$. XGBoost construit une somme de K fonctions f_k appartenant à l'espace des arbres de décision $\mathcal{F}$:

$$\hat{y}_i = \sum_{k=1}^{K} f_k\,(x_k), \text{ où } f_k \in \mathcal{F}$$

L'objectif de l'algorithme est alors de **minimiser une fonction objectif régularisée**, composée de deux termes :

$$\Im(\theta) = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) \; + \; \sum_{k=1}^{K} \Omega(f_k)$$

Où θ represente l'ensemble des **paramètres du modèle à optimiser (**structure des arbres de décision, les poids associés aux feuilles et les hyperparamètres**)**

• Le **premier terme** $\ell(.,.)$ mesure l'erreur entre la prédiction et la valeur réelle (via la **log-loss** dans le contexte de la classification).

• Le **second terme** $\Omega(f_k)$ agit comme pénalité sur la complexité du modèle. Il est défini par :

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2$$

où **T** est le nombre de feuilles de l'arbre, **wj** les scores associés aux feuilles, et **γ,λ** sont des coefficients de régularisation.

L'ajout de chaque nouvel arbre dans l'ensemble se fait **de manière incrémentale**, en ajustant les prédictions précédentes :

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$$

Chaque nouvel arbre f_t est construit de façon à **minimiser une approximation au second ordre** de la fonction de perte. Cela permet une optimisation plus efficace, comme détaillé dans les travaux de Chen et Guestrin.

Par ailleurs, pour améliorer la capacité de généralisation du modèle, XGBoost intègre :

- une **stratégie de "shrinkage"** (apprentissage par pas), contrôlée par un **taux d'apprentissage** η,
- un **sous-échantillonnage des observations** (row subsampling),
- et un **sous-échantillonnage des variables** (column subsampling), inspirés des méthodes de Random Forest.

Ces mécanismes permettent à XGBoost de trouver un bon compromis entre **biais** et **variance**, tout en conservant une excellente performance sur les données tabulaires.

4.2 ANALYSE DE SURVIE : Modèles de survie et prédiction dans le temps

Introduction

L'analyse de survie, également appelée « analyse de durée » (ou *time-to-event analysis* en anglais), constitue un cadre statistique essentiel pour modéliser le temps écoulé avant la survenue d'un événement d'intérêt. Initialement développée en biométrie pour étudier la durée de survie des patients, cette approche s'est étendue à de nombreux domaines, y compris le marketing, la finance, l'industrie et, dans le cadre de ce travail, l'e-commerce. Dans notre étude, l'événement d'intérêt est le **churn client**, c'est-à-dire l'instant où un client cesse d'acheter ou interagit pour la dernière fois avec la plateforme.

Ce paradigme est particulièrement adapté à la modélisation du **comportement dynamique des clients dans le temps**, permettant d'aller au-delà des simples approches prédictives binaires (churn / non churn) pour intégrer la **composante temporelle**. L'objectif est alors de **prédire la probabilité de survie d'un client à différents horizons temporels**, de comparer les durées de

survie entre groupes, ou encore d'estimer les effets de variables explicatives sur le risque de churn.

L'analyse de survie offre aussi la possibilité de gérer **les données censurées**, c'est-à-dire les clients pour lesquels le churn n'a pas encore été observé à la date d'analyse. Cette caractéristique est particulièrement précieuse dans les contextes transactionnels, où une proportion significative de clients est encore active.

4.2.1 Base de l'analyse de survie : Fonction de survie et fonction de hasard

La base conceptuelle de l'analyse de survie repose sur plusieurs fonctions fondamentales. La première est la **fonction de survie** S(t)S(t)S(t), définie comme la probabilité qu'un individu ou une entité "survive" au-delà du temps ttt, c'est-à-dire que l'événement d'intérêt ne se soit pas encore produit à ce moment-là.

Mathématiquement :

$$S(t) = \mathbb{P}(T > t)$$

où T est une variable aléatoire représentant le temps jusqu'à l'événement.

Cette fonction est décroissante, avec S(0) = 1 et $\lim_{t \to +\infty} S(t) = 0$ sous des conditions classiques.

- **Fonction de densité de probabilité** f(t):
  Elle désigne le taux instantané de survenue de l'événement à l'instant ttt, sachant que l'individu a "survécu" jusque-là. Elle est définie comme :

$$f(t) = \frac{d}{dt}[1 - S(t)] = -\frac{d}{dt}S(t)$$

Fonction de risque (ou hasard) h(t) :
Cette fonction désigne le taux instantané de survenue de l'événement à l'instant t, sachant que l'individu a "survécu" jusque-là.

Elle se définit par :

$$h(t) = \lim_{\Delta t \to +\infty} \frac{\mathbb{P}\,(t < T \leq t + \Delta t \mid T > t)}{\Delta t}$$

Elle n'est pas une probabilité mais une intensité : elle représente la « vitesse » à laquelle les individus churnent à l'instant t, parmi ceux encore présents à ce moment.

La fonction de hasard peut varier au cours du temps, ce qui permet de capter la dynamique temporelle du risque.

Par ailleurs, on introduit souvent la **fonction de risque cumulée** H(t), définie comme :

$$H(t) = \int_0^t h(u)du$$

Le lien entre ces fonctions permet également de reconstruire S(t) à partir de h(t), via la relation suivante :

$$S(t) = \exp\left(-\int_0^t h(u)du\right)$$

Cette expression montre que la survie dépend cumulativement du risque au cours du temps.

Les modèles paramétriques de survie supposent une forme spécifique pour h(t) (ex. : exponentielle, Weibull), tandis que les modèles semi-paramétriques, comme le **modèle de Cox**, introduisent une structure multiplicative sur le risque en fonction de variables explicatives, sans imposer de forme à la fonction de risque de base. Nous détaillons ce modèle dans la section suivante.

Ces fonctions constituent le socle de tous les modèles de survie. Dans les sections suivantes, elles seront utilisées pour estimer la survie empirique (Kaplan-Meier), ajuster des modèles paramétriques (Weibull AFT), semi-paramétriques (Cox PH) ou encore neuronaux (DeepSurv, DeepHit), et évaluer l'effet des caractéristiques clients sur la durée de leur engagement.

4.2.2 L'estimateur de Kaplan-Meier

L'estimateur de **Kaplan–Meier** constitue l'un des outils statistiques les plus couramment utilisés pour modéliser les fonctions de survie, en particulier dans les contextes où aucune hypothèse paramétrique n'est posée sur la forme de la distribution des durées [25]. Il permet d'estimer la probabilité de **non-occurrence d'un événement (ici, le churn)** à un instant donné, en tenant compte de la censure (clients encore actifs).

Contrairement aux estimateurs paramétriques classiques, le Kaplan–Meier s'est avéré être **non biaisé** dans la majorité des cas pratiques, avec des **pertes d'efficacité négligeables**, sauf dans des situations extrêmes comme les très petits échantillons [26].

Définition formelle :
Soient $t_1<t_2<\cdots<t_k$ les instants où un événement (ici, un churn) est observé. À chaque instant t_j, on note :

- d_j : le nombre d'événements (churns) survenus à t_j,
- n_j : le nombre d'individus encore en observation juste avant tjt_jtj (clients « à risque »),
- c_j : le nombre d'observations censurées à t_j (si applicable).

L'estimateur de Kaplan-Meier s'écrit :

$$\hat{S}(t) = \prod_{tj \le t}\left(1 - \frac{d_i}{n_j}\right)$$

Chaque terme correspond à la probabilité conditionnelle de ne pas churner à t_j, sachant que le client est toujours actif juste avant. Cette construction assure que l'estimateur est **étagé** (en escalier), avec des discontinuités aux instants d'événement.

Le résultat produit est une **courbe en escalier**, où chaque saut correspond à un événement (ici, un churn), souvent représentée avec des **intervalles de confiance** autour de la courbe empirique [19-4].

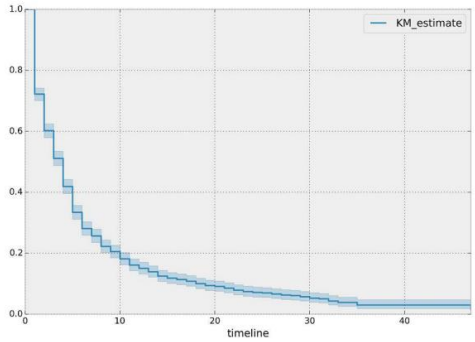


Figure : Courbe de survie de Kaplan-Meier générée à partir du jeu de données *lifelines* sur les démocraties et dictatures, avec intervalles de confiance [19-4]

Interprétation :
La courbe $\hat{S}(t)$ estimée trace, pour chaque instant t, la probabilité qu'un client reste actif au-delà de ce temps. Elle permet d'observer visuellement :

- la vitesse de décroissance de la rétention client ;
- les seuils temporels critiques (ex. : 6 mois après l'acquisition) ;
- l'effet de certaines covariables (ex. canal d'acquisition) sur la durée de vie client, via la comparaison de courbes par sous-groupes.

Cependant, cette approche présente une **limite importante** : elle **n'intègre pas les covariables individuelles** (démographiques, comportementales, etc.). Elle est donc idéale pour visualiser la durée de vie globale dans un groupe, mais ne permet pas de comparer formellement les effets de différentes variables explicatives. Il est néanmoins possible de segmenter l'échantillon selon quelques variables catégorielles pour construire plusieurs courbes, facilitant les comparaisons exploratoires [25,26-4].

4.2.3 Modèle de Cox à risques proportionnels

Le **modèle de Cox à risques proportionnels**, introduit par Sir David Cox en 1972, constitue l'un des piliers de l'analyse de survie moderne, en particulier dans des contextes où la forme exacte de la fonction de risque n'est pas connue a priori. Ce modèle **semi-paramétrique** permet de quantifier l'effet de variables explicatives (ou covariables) sur le risque de survenue de l'événement, tout en laissant la fonction de risque de base $h_0(t)$ non spécifiée.

a) Formulation mathématique

Le modèle de Cox suppose que le risque instantané pour un individu i, de vecteur de covariables $x_i=(x_{i1},x_{i2},...,x_{ip})$, s'écrit :

$$h(t \mid x_i) = h_0(t) \cdot \exp(x_i^\top \beta)$$

où:

- $h_0(t)$ est la fonction de risque de base (commune à tous les individus),
- $\beta=(\beta_1,...,\beta_p)$ est le vecteur des coefficients à estimer.

Ce modèle repose sur **l'hypothèse des risques proportionnels** : le ratio des fonctions de risque de deux individus quelconques reste constant dans le temps, soit :

$$\frac{h(t \mid x_1)}{h(t \mid x_2)} = \exp((x_1 - x_2)^\top \beta)$$

Ainsi, les covariables influencent le niveau de risque, mais non sa dynamique temporelle.

b) Estimation des paramètres

L'estimation du vecteur β s'effectue par **maximisation de la vraisemblance partielle**, une approche ingénieuse introduite par Cox permettant d'éviter la spécification de $h_0(t)$. Cette méthode repose uniquement sur l'ordre des temps de survenue et exploite la nature censurée des données.

La vraisemblance partielle s'écrit, pour n individus dont les événements sont observés aux temps $t_1<t_2<\cdots<t_k$, comme :

$$\mathcal{L}(\beta) = \prod_{t=1}^{k} \frac{\exp(x_i^\top \beta)}{\sum_{j \in \mathcal{R}(t_i)} \exp(x_j^\top \beta)}$$

où $\mathcal{R}(t_i)$ désigne le **set de risque** au temps t_i, c'est-à-dire les individus encore à risque à cet instant.

c) Interprétation des coefficients

Chaque coefficient β_j peut être interprété via le **hazard ratio** (HR) associé à la variable x_j :

$$HR = \exp(\beta_j)$$

Un HR supérieur à 1 indique que l'augmentation de xjx_jxj accroît le risque de survenue de l'événement, tandis qu'un HR inférieur à 1 traduit un effet protecteur.

d) Vérification de l'hypothèse des risques proportionnels

L'hypothèse peut être testée à l'aide :

- des **résidus de Schoenfeld**, qui doivent être indépendants du temps,
- du **test global de Schoenfeld**, pour vérifier si l'ensemble des covariables respecte cette hypothèse.

En cas de violation, plusieurs stratégies sont envisageables, telles que l'inclusion d'interactions avec le temps ou le recours à des modèles alternatifs comme le modèle présenté ci-dessous.

4.2.4 Le modèle de survie Weibull AFT

a) Positionnement du modèle AFT

Les modèles **Accelerated Failure Time (AFT)** constituent une famille **paramétrique** de modèles de survie où les covariables agissent **directement sur l'échelle du temps** : elles "accélèrent" ou "ralentissent" le temps jusqu'à l'événement. Contrairement au **modèle de Cox à risques proportionnels (PH)**, qui impose un effet multiplicatif des covariables sur le **risque** (hazard), le modèle AFT impose un effet multiplicatif sur le **temps de survie**. Cette différence entraîne des interprétations particulièrement intuitives en contexte managérial : un coefficient positif peut être lu comme une **augmentation (multiplicative) du temps attendu avant churn** (time ratio > 1), et inversement.

Parmi les distributions paramétriques utilisables dans un cadre AFT (exponentielle, log-normale, log-logistique, Gompertz, etc.), **la loi de Weibull occupe une place singulière** : elle est **la seule** à pouvoir être écrite **à la fois** sous forme **AFT** *et* **PH** (proportionnalité des risques), ce qui en fait un pont conceptuel entre les deux grandes familles de modèles de survie. Cette dualité confère au Weibull une grande flexibilité pratique et théorique. [27-4, 28-4, 29-4]

b) La Distribution de Weibull

La distribution de Weibull est largement utilisée pour modéliser les temps de survie en raison de sa capacité à capturer des dynamiques de risque variées. Soit T le temps jusqu'à l'événement (par exemple, le churn). La fonction de densité de probabilité d'une variable suivant une distribution de Weibull à deux paramètres est :

$$f(t|\,\gamma,\rho) = \frac{\gamma}{\rho}\left(\frac{t}{\rho}\right)^{\gamma-1} \exp\left(-\left(\frac{t}{\rho}\right)^{\gamma}\right) \quad (1)$$

où $\gamma > 0$ est le paramètre de forme et $\rho > 0$ est le paramètre d'échelle. La fonction de survie S(t), représentant la probabilité que l'événement ne se produise pas avant t, est :

$$S(t) = \exp\left(-\left(\frac{t}{\rho}\right)^{\gamma}\right). \qquad (2)$$

La fonction de risque h(t), mesurant la probabilité instantanée de l'événement à t, est donnée par :

$$h(t) = \frac{\gamma}{\rho}\left(\frac{t}{\rho}\right)^{\gamma-1} \qquad (3)$$

Le paramètre de forme γ détermine la dynamique du risque :

- γ = 1 : Risque constant, équivalent à une distribution exponentielle.
- γ > 1 : Risque croissant, typique des phénomènes d'usure ou dedégradation.
- γ < 1 : Risque décroissant, caractéristique des sorties précoces suivies d'une stabilisation.

c) Formulation du Modèle Weibull AFT avec covariables

Dans le cadre du modèle *Accelerated Failure Time* (AFT), le temps de survie *Ti* pour un individu *i* est modélisé sur une échelle logarithmique pour capturer l'effet des covariables sur la durée jusqu'à l'événement (le churn dans un contexte de cette étude). La formulation générale du modèle AFT est donnée par :

$$\log T_i = \mu + \boldsymbol{x_i^{\intercal}\beta} + \sigma\varepsilon_i \quad (1)$$

où :

• $\mathbf{x}_i$ est le vecteur des covariables pour l'individu *i*,
• $\boldsymbol{\beta}$ est le vecteur des coefficients associés aux covariables,
• μ est l'intercept,
• $\sigma > 0$ est le paramètre d'échelle,
• ε_i est une erreur aléatoire suivant une distribution de Gumbel (valeur extrême de type I), ce qui implique que T_i suit une distribution de Weibull.

Dans le cas du modèle Weibull AFT, le paramètre d'échelle individuel $\rho(\mathbf{x}_i)$ est défini comme :

$$\rho(\boldsymbol{x_i}) = \rho_0 \, exp(\boldsymbol{x_i^{\intercal}\,\beta}) \quad , (2)$$

où $\rho_0 = \exp(\mu)$ représente l'échelle de base. La fonction de survie conditionnelle pour un individu avec covariables $\mathbf{x}i$ s'exprime alors comme :

$$S(t|x_i) = \exp\left(-\left(\frac{t}{\rho(x_i)}\right)^{\gamma}\right)$$

où $\gamma > 0$ est le paramètre de forme, commun à tous les individus, qui contrôle la dynamique du risque (croissant, décroissant ou constant). La fonction de risque correspondante est :

$$h(t \mid x_i) = \frac{\gamma}{\rho(x_i)}\left(\frac{t}{\rho(x_i)}\right)^{\gamma-1}$$

On montre (en particulier dans les implémentations comme **lifelines**) que l'**acceleration factor** (ou **time ratio**) associé à une covariable x_j vaut :

$$TR_j = \exp(\beta_j)$$

ce qui signifie qu'une variation d'une unité de x_j **multiplie le temps caractéristique** (médian, par exemple) par exp (β_j).
Ainsi :

- $TR_j > 1$ (i.e $\beta_j < 0$): l'événement (churn) est **ralenti** (le client "survit" plus longtemps),
- $TR_j < 1$: l'événement est **accéléré**.

Lien avec le modèle PH (proportionnalité des risques).
Pour la loi de Weibull, il existe une reparamétrisation telle que le même modèle puisse être écrit en **PH** :

$$h(t\,|x) = h_0(t) \times \exp\,(x^{\intercal}\tilde{\beta})$$

où $\tilde{\beta} = -\gamma\beta$ (relation de correspondance sous certaines conventions de paramétrage). Cette propriété *unique* justifie l'usage du Weibull comme modèle "pivot" entre AFT et PH. [*27-4, 28-4, 30-4]*

d) Estimation par maximum de vraisemblance sous censure

Les paramètres (β,ρ,γ) sont estimés via le **maximum de vraisemblance** en tenant compte de la **censure à droite** (la plus fréquente en pratique du churn) *[39-4]* :

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} [f(t_i|x_i)]^{\delta_i}\ [S(t_i|x_i)]^{1-\delta_i}$$

où :

- t_i est le temps observé (churn ou censure),
- δ_i=1 si l'événement (churn) est observé, 0 sinon (censuré),
- $f(\cdot|x_i)$ et $S(\cdot|x_i)$ sont la densité et la survie paramétriques conditionnelles,
- θ= (β,ρ,γ)

L'optimisation est réalisée numériquement (Newton-Raphson, quasi-Newton, etc.). Les implémentations modernes (Python `lifelines.WeibullAFTFitter` [35-4]) fournissent directement les estimateurs, leurs écarts-types, intervalles de confiance et tests de Wald/LR.

e) Diagnostics, validation et adéquation au Weibull

Pour valider l'adéquation du modèle Weibull AFT, plusieurs outils sont utilisés :

- Transformation log-log : Le tracé de log(- log S(t)) contre logt doit être linéaire, comme implémenté dans le code avec lifelines. Cette linéarité confirme l'adéquation de la distribution de Weibull. [36-4, 29-4]
- Critères d'information : Les critères AIC et BIC comparent le modèle Weibull AFT à d'autres distributions (log-normale, log-logistique) ou au modèle de Cox. [37-4, 29-4]

- Résidus : Les résidus de Cox-Snell ou de déviance évaluent la qualité d'ajustement. Une superposition des survies observées et prédites par quantiles de risque peut également être réalisée. [38-4]

4.2.5 Évaluation des performances des modèles de survie : le Brier Score et le C-index

L'évaluation des performances prédictives constitue une étape essentielle de toute analyse de survie rigoureuse. Elle ne se limite pas à une simple comparaison statistique : elle permet de juger de la pertinence scientifique des modèles tout en garantissant leur utilité pratique dans des contextes appliqués, tels que la prédiction du churn client en e-commerce. Cette évaluation est rendue particulièrement complexe par la **présence de censure**, qui distingue fondamentalement l'analyse de survie des modèles classiques de régression ou de classification. Dans ce cadre, deux métriques complémentaires et largement validées par la littérature sont mobilisées : **le Brier Score** et le **Concordance Index** (C-index). Leur combinaison offre une appréciation robuste et équilibrée des performances, en couvrant deux dimensions fondamentales de la qualité prédictive : le calibrage et la discrimination.

Le Brier Score corrigé par IPCW

Le Brier Score, introduit par Brier (1950), mesure l'écart quadratique moyen entre les probabilités prédictives d'un modèle et les événements réellement observés. Appliqué à la survie, il évalue la qualité du calibrage des probabilités, c'est-à-dire la correspondance entre les prédictions et les résultats empiriques. Un score faible traduit donc une bonne adéquation des prévisions probabilistes à la réalité.

Cependant, l'utilisation directe de cette métrique est biaisée en présence de censure, caractéristique incontournable des données de survie. Pour un individu censuré, le moment exact de l'événement n'est pas observé, ce qui rend l'évaluation du décalage entre prédiction et réalité impossible. Pour corriger ce biais, Graf et al. (1999) puis Gerds & Schumacher (2006) ont proposé une version pondérée par l'**Inverse Probability of Censoring Weighting (IPCW)**. Cette méthode repose sur une estimation de Kaplan-Meier de la censure et attribue à chaque individu un poids reflétant sa probabilité d'être encore observable à un instant donné.

Ainsi, le Brier Score corrigé IPCW est aujourd'hui considéré comme une référence incontournable dans l'évaluation des modèles de survie, car il :

- garantit une comparaison équitable entre modèles, indépendamment du degré de censure,
- renseigne sur le calibrage des probabilités à différents horizons temporels,
- assure la robustesse des conclusions dans des environnements appliqués, comme le e-commerce, où de nombreux clients restent actifs (et donc censurés) au moment de l'analyse.

Le Concordance Index (C-index)

Alors que le Brier Score renseigne sur la précision absolue des probabilités prédites (*calibrage*), le Concordance Index (C-index) mesure la capacité relative du modèle à ordonner correctement

les individus selon leur risque (*discrimination*). Proposé par Harrell et al. (1982), il est souvent rapproché de l'aire sous la courbe ROC (AUC) utilisée en classification binaire. Le C-index mesure la probabilité qu'un modèle attribue un risque plus élevé à un individu qui expérimente l'événement avant un autre, resté plus longtemps en observation.

Formellement, il correspond à la proportion de paires concordantes parmi l'ensemble des paires comparables (c'est-à-dire celles pour lesquelles l'ordre des événements est identifiable). Sa valeur varie entre :

- 0,5, correspondant à une performance aléatoire,
- 1, représentant une discrimination parfaite.

Un C-index élevé traduit donc une hiérarchisation correcte des risques, qualité essentielle lorsqu'il s'agit, par exemple, d'identifier les clients les plus susceptibles de quitter un service dans un horizon donné.

Le C-index présente plusieurs atouts majeurs :

- il est non paramétrique, ne reposant sur aucune hypothèse de distribution des temps de survie,
- il est naturellement robuste à la censure, car seules les paires comparables sont prises en compte,
- il offre une mesure intuitive de la discrimination, souvent plus parlante pour des praticiens ou décideurs que des indicateurs purement techniques.

Il convient toutefois de souligner une limite : lorsque la censure n'est pas aléatoire mais dépendante du risque (censure informative), le C-index peut être biaisé. Ce point renforce l'importance d'utiliser cette métrique en complément d'autres mesures.

Complémentarité des deux métriques

Pris isolément, ni le C-index ni le Brier Score ne suffisent à évaluer la performance globale d'un modèle de survie :

- un modèle peut très bien discriminer les individus (C-index élevé) mais fournir des probabilités mal calibrées,
- ou inversement, avoir un bon calibrage (Brier Score faible) mais échouer à hiérarchiser correctement les risques.

C'est pourquoi la combinaison des deux approches est aujourd'hui considérée comme une pratique méthodologique standard (Graf et al., 1999 ; Gerds & Schumacher, 2006 ; Harrell et al., 1982). Cette complémentarité assure une évaluation complète :

- le Brier Score IPCW rend compte du calibrage probabiliste,
- le C-index évalue la discrimination entre individus.

Dans ce mémoire, ces deux métriques constituent un socle théorique et méthodologique essentiel pour comparer la pertinence de différents modèles de survie — qu'ils soient semi-paramétriques (Cox proportionnel), paramétriques (Weibull AFT) ou neuronaux (DeepSurv,

DeepHit). Cette approche duale permet de garantir non seulement la validité statistique des modèles, mais aussi leur pertinence opérationnelle dans un système de fidélisation en e-commerce, où l'identification fiable et exploitable des clients à risque de churn est stratégique.

4.3. Prédiction de valeur vie client

4.3.1 Le Modèle BG/NBD

a. Principes et Hypothèses Fondamentales

Le modèle **BG/NBD (Beta-Geometric/Negative Binomial Distribution)** constitue une extension du modèle classique **Pareto/NBD**, mais il introduit une différence conceptuelle clé :

- dans le Pareto/NBD, l'attrition peut se produire à tout moment, indépendamment des achats ;
- dans le BG/NBD, l'attrition ne peut survenir qu'immédiatement après une transaction.

Ce modèle repose sur six hypothèses [4-10X]:

- Le client traverse deux étapes au cours de sa relation avec une entreprise donnée **:**
  - être actif pendant une certaine période de temps
  - devenir définitivement inactif

- tant qu'il est actif, le nombre de transactions effectuées par un client suit un processus de Poisson avec un taux de transaction λ. Le temps entre les transactions suit une distribution exponentielle :

$$f\left(t_j \middle| t_{j-1}; \lambda\right) = \lambda e^{-\lambda(t_j - t_{j-1})}, \qquad t_j > t_{j-1} \geq 0$$

- l'hétérogénéité du taux de transaction λ entre les clients suit une distribution gamma :

$$f(\lambda \mid r, \alpha) = \frac{\alpha^r \lambda^{r-1} e^{-\lambda\alpha}}{\Gamma(r)}, \quad \lambda > 0$$

- Après chaque transaction, un client devient inactif avec une probabilité p. Le moment d'abandon suit une distribution géométrique décalée :

$$P(inactif\ immédiatement\ après\ la\ j^{ème}\ transaction) = p(1-p)^{j-1},\ j = 1,2,3,\ldots$$

- L'hétérogénéité de la probabilité d'abandon p suit une distribution bêta :

$$\mathrm{f}\,(p \mid a, b) = \frac{p^{a-1}(1-p)^{b-1}}{\beta(a,b)}, \quad 0 \leq p \leq 1$$

où $\beta(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ est la fonction bêta.

- Le taux de transaction λ et la probabilité d'abandon p varient indépendamment entre les clients.

b. Variables d'Entrée : Récence, Fréquence et Âge Client

Le modèle est opérationnalisé à partir de trois caractéristiques fondamentales de chaque client [4-10X] :

- **Recency (R)** : différence (en unités de temps) entre la première et la dernière transaction observée.
- **Frequency (F)** : nombre de transactions répétées pendant la période d'observation, i.e. F=nombre total de transactions−1.
- **Time (T)** : durée (en unités de temps) entre la première transaction et la fin de la période d'observation.

Ces trois variables résument le comportement transactionnel historique d'un client et permettent de dériver :

- la probabilité qu'un client soit encore « vivant » (actif),
- le nombre attendu de transactions futures dans une période donnée,
- la distribution prédictive du nombre d'achats.

c. Développements mathématiques

a- Développement au Niveau Individuel

Fonction de Vraisemblance

Pour un client ayant effectué x transactions dans la période (0, T] avec la dernière transaction à t_x, la fonction de vraisemblance individuelle s'écrit :

$$\mathcal{L}(\lambda, p \mid X = x, t_x, T) = (1-p)^x \lambda^x e^{-\lambda T} + \delta_{x>0} \cdot p(1-p)^{x-1} \lambda^x e^{-\lambda t x}$$

Où $\delta_{x>0} = 1$ si x>0, 0 sinon.

Cette expression capture deux scenarios : le client reste actif avec une probabilité $(1-p)^x$ ou devient inactif après la $x^{ème}$ transaction avec une probabilité $p(1-p)^{x-1}$

Distribution prédictive

La probabilité d'observer exactement x transactions dans une fenêtre de longueur t, conditionnelle aux paramètres individuels, est donnée par :

$$P(X(t) = x \mid \lambda, p) = (1-p)^x \frac{(\lambda t)^x e^{-\lambda t}}{x!} + \delta_{x>0} \cdot p(1-p)^{x-1} \left[1 - e^{-\lambda t} \sum_{j=0}^{x-1} \frac{(\lambda t)^j}{j!} \right]$$

Nombre Attendu de Transactions

Le nombre attendu de transactions dans un intervalle de longueur t pour un client avec paramètres (λ, p) est :

$$E[X(t) \mid \lambda, p] = \frac{1}{p} - \frac{1}{p} e^{-\lambda p t}$$

Cette expression reflète le fait que le processus suit initialement un taux λ mais est "atténué" par la probabilité d'abandon p.

b- Développement pour un Client Aléatoire

Jusqu'ici, nous avons travaillé au niveau individuel, en conditionnant sur des valeurs fixes du taux de transaction λ et de la probabilité d'abandon p. Pour obtenir des formules applicables à un client choisi **au hasard dans la population**, il est nécessaire d'intégrer ces résultats sur les distributions a priori de λ et p.

Paramétrisation des lois a priori

Le modèle **BG/NBD** introduit deux sources d'hétérogénéité inter-clients :

- **Taux de transaction λ\lambdaλ**
  On suppose que le taux de transaction de chaque client suit une loi Gamma :

$$\lambda \sim \text{Gamma}(r,\alpha)$$

où :

- r>0: paramètre de forme, reflétant la régularité moyenne des transactions,
- α>0 : paramètre d'échelle contrôlant la dispersion.

- **Probabilité d'abandon p**
  On suppose que la probabilité d'attrition suit une loi Beta :

$$p \sim \text{Beta}(a,b)$$

où :

- a>0 : premier paramètre de forme,
- b>0 : second paramètre de forme.

Fonction de Vraisemblance Agrégée

Pour un client ayant effectué x transactions dans la fenêtre d'observation (0,T], avec dernière transaction à t_x, la vraisemblance intégrée sur λ et p est :

$$\mathcal{L}(r, \alpha, a, b \mid X = x, t_x, T) = \frac{\beta(a, b+x)}{\beta(a,b)} \cdot \frac{\Gamma(r+x)\alpha^r}{\Gamma(r)(\alpha+T)^{r+x}} + \delta_{x>0} \cdot \frac{\beta(a+1, b+x-1)}{\beta(a,b)} \cdot \frac{\Gamma(r+x)\alpha^r}{\Gamma(r)(\alpha+t_x)^{r+x}}$$

Distribution Prédictive

La probabilité d'observer x transactions pendant un horizon ttt, pour un client tiré au hasard de la population.

$$P(X(t)=x \mid r,\alpha,a,b) = \frac{\beta(a,b+x)}{\beta(a,b)} \cdot \frac{\Gamma(r+x)}{\Gamma(r)x!}\left(\frac{\alpha}{\alpha+t}\right)^r \left(\frac{t}{\alpha+t}\right)^x + \delta_{x>0} \cdot \frac{\beta(a+1,b+x-1)}{\beta(a,b)}\left[1-\left(\frac{\alpha}{\alpha+t}\right)^r \sum_{j=0}^{x-1} \frac{\Gamma(r+j)}{\Gamma(r)j!}\left(\frac{t}{\alpha+t}\right)^j\right].$$

Nombre Attendu de Transactions Futures

Pour un client avec un historique d'achat (X=x,t_x,T), le nombre attendu de transactions dans une période future de longueur t est :

$$E[Y(t) \mid X=x,t_x,T,r,\alpha,a,b] = \frac{\frac{a+b+x-1}{a-1}\left[1-\left(\frac{\alpha+T}{\alpha+T+t}\right)^{r+x} + {}_2F_1(r+x,b+x;a+b+x-1;t\alpha+T+t)\right]}{1+\delta_{x>0}\frac{a}{b+x-1}\left(\frac{\alpha+T}{\alpha+t_x}\right)^{r+x}}$$

où ${}_2F_1(\cdot)$ désigne la fonction hypergéométrique gaussienne et Y(t) est le nombre de transactions futures.

Estimation des Paramètres

Les paramètres (r,α,a,b) sont estimés par maximum de vraisemblance en maximisant la log-vraisemblance agrégée :

$$\mathcal{LL}(r,\alpha,a,b) = \sum_{i=1}^{N} ln\left[\mathcal{L}(r,\alpha,a,b \mid X_i=x_i,t_{x_i},T_i)\right]$$

Pour davantage de détails théoriques et méthodologiques sur le modèle BG/NBD et ses fondements, le lecteur pourra se référer aux travaux de Fader, Hardie & Lee *[4-X1]* ainsi que de Schmittlein, Morrison & Colombo *[4-X2]*.

d. Utilité et Applications

Le BG/NBD fournit des prédictions fiables pour :

- la probabilité qu'un client soit encore actif,
- le nombre attendu de transactions futures dans un horizon donné,
- la segmentation des clients sur la base de leurs comportements transactionnels (notamment via RFM).

Ce modèle a été appliqué avec succès dans divers secteurs, tels que la grande distribution *[4-12X]* ou la banque de détail en Afrique du Nord *[4-12X]*, où il a permis une segmentation fine et des prévisions exploitables pour le calcul de la **Customer Lifetime Value (CLV)**.

Cependant, le BG/NBD ne prédit pas directement la valeur monétaire des transactions. Pour cette raison, il est couramment couplé au modèle **Gamma-Gamma**, qui estime la dépense moyenne par transaction à partir de la distribution monétaire observée [*14*]. L'association BG/NBD + Gamma-Gamma constitue ainsi l'approche probabiliste standard pour l'évaluation du CLV.

4.3.2 Modèle Gamma-Gamma

Alors que le modèle **BG/NBD** permet de prédire la fréquence future des transactions des clients en tenant compte de la probabilité de churn, il ne fournit aucune information directe sur la valeur monétaire associée à chaque transaction.

Pour compléter cette analyse, **le modèle Gamma-Gamma (Fader et al., 2005)** est introduit afin d'estimer la dépense monétaire moyenne d'un client actif. Ce modèle constitue une étape essentielle dans le calcul de la **Customer Lifetime Value (CLV)**, car il permet de relier la fréquence des achats à leur valeur financière.

a. Hypothèses du modèle

Le modèle Gamma-Gamma repose sur un ensemble d'hypothèses structurelles :

- **Indépendance fréquence–montant** :
  La fréquence des transactions d'un client est supposée **indépendante** du montant dépensé par transaction. Ainsi, un client qui achète plus souvent n'est pas nécessairement celui qui dépense plus à chaque achat.
- **Valeur monétaire constante par client** :
  Chaque client iii possède une **valeur monétaire moyenne latente** M_i, qui reste constante dans le temps (il n'y a pas de tendance à la hausse ou à la baisse des montants pour un client donné).
- **Hétérogénéité entre clients** :
  Les clients diffèrent dans leurs dépenses moyennes, et cette hétérogénéité est modélisée par une **loi Gamma**.

b. Variables et données d'entrée

Pour chaque client, les données nécessaires sont :

- x: le nombre total de transactions observées.
- m_x : la moyenne empirique du montant des transactions historiques d'un client :

$$m_x = \frac{1}{x}\sum_{i=1}^{x} z_i$$

où z_i est la valeur monétaire de la i-ème transaction du client.

c. Formulation mathématique

Le modèle repose sur une structure hiérarchique :

1. La dépense moyenne par client est notée M.
2. Conditionnellement à M, chaque transaction suit une distribution Gamma.
3. L'hétérogénéité des M dans la population suit également une loi Gamma, d'où le nom **Gamma-Gamma**.

Ainsi, la distribution a posteriori de M donnée l'historique des dépenses (Farder & Hardie) est :

$$E[M \mid p, q, \gamma, m_x, x] = \left(\frac{q-1}{px+q-1}\right)\frac{p\gamma}{q-1} + \frac{px}{px+q-1}\, m_x$$

où :

- p : paramètre de forme de la Gamma pour les transactions individuelles,
- q: paramètre de forme pour la distribution de M (dépense moyenne par client),
- γ: paramètre d'échelle pour la distribution de M.

d. Estimation des paramètres

Les paramètres (p,q,γ) sont estimés à partir des données de transactions par **maximum de vraisemblance (MLE)** ou par **méthodes bayésiennes** (selon la littérature).

Une fois les paramètres estimés, l'espérance conditionnelle $E(M|m_x,x)$ fournit la **valeur monétaire moyenne attendue** d'un client donné, ajustée à son comportement observé.

4.3.3 Utilisation jointe des modèles BG/NBD et Gamma-Gamma pour l'estimation de la CLV

Pris séparément, le modèle **BG/NBD** et le modèle **Gamma-Gamma** n'offrent qu'une vision partielle du comportement client :

- le **BG/NBD** permet de prédire la **fréquence d'achat future** en tenant compte de la probabilité de churn,
- le **Gamma-Gamma** estime la **valeur monétaire moyenne attendue par transaction**.

Cependant, la **Customer Lifetime Value (CLV)** est une métrique qui combine ces deux dimensions :

$$CLV = Fréquence\ attendue\ des\ transactions\ \times Valeur\ monétaire\ moyenne\ attendue$$

Ainsi, l'utilisation conjointe des deux modèles fournit une approche cohérente et robuste pour estimer la CLV, comme l'ont proposé **Fader et Hardie** et repris dans de nombreuses applications en marketing quantitatif.

Le modèle **BG/NBD** capture l'hétérogénéité des taux de transactions et des probabilités de churn. Il fournit $E[Y_i(t)]$, le **nombre attendu de transactions futures** sur un horizon t pour un client i.

Le modèle **Gamma-Gamma** capture l'hétérogénéité des montants moyens de transaction entre clients. Il fournit $E[M_i|m_x,x]$, la **valeur monétaire moyenne attendue** des transactions d'un client, conditionnellement à ses dépenses passées.

Ainsi, en combinant les deux résultats, la CLV pour un client i sur un horizon temporel ttt peut s'écrire :

$$\mathrm{CLV}_i(t) \;=\; E[Y_i(t) \mid r, \alpha, a, b] \;\times\; E[M_i \mid p, q, \gamma, m_x, x]$$

où :

- $E[Y_i(t)]$ est issu du **BG/NBD**,
- $E[M_i]$ est issu du **Gamma-Gamma**,
- (r,α,a,b) sont les paramètres du BG/NBD,
- (p,q,γ) sont les paramètres du Gamma-Gamma.

XConclusion

Chapitre 5 – Implémentation du pipeline analytique et résultats

Ce chapitre constitue la mise en pratique rigoureuse de l'approche méthodologique. Il met en œuvre les techniques présentées précédemment sur les données, en suivant une logique scientifique reproductible.

5.1 Présentation du jeu de données

5.1.1 Description générale du jeu de données

Le présent mémoire s'appuie sur un jeu de données synthétique, mais réaliste, simulant le fonctionnement d'une plateforme e-commerce B2C à grande échelle. Ce dataset comprend **5630 observations** réparties sur **19 variables explicatives**, et une variable cible `Churn` définissant si un client a quitté la plateforme ou y est resté actif. Chaque ligne représente un client unique, décrit selon ses caractéristiques sociodémographiques, comportementales, transactionnelles et digitales. La richesse de cette base permet d'aborder la problématique du **churn** sous un angle à la fois descriptif, prédictif et temporel.

5.1.2 Nature et signification des variables

Une lecture attentive des attributs révèle une forte hétérogénéité des informations disponibles, réparties comme suit :

- **Identifiant unique du client (`CustomerID`)** : Clé primaire assurant l'unicité de chaque individu, indispensable pour les opérations de fusion, de traçabilité et de segmentation.

- **Variable cible (`Churn`)** : Indicateur binaire représentant le départ effectif du client. Cette colonne constitue le cœur de l'analyse prédictive.
- **Ancienneté (`Tenure`)** : Nombre de mois depuis lesquels le client est actif sur la plateforme. Elle reflète la stabilité de la relation client.
- **Appareil préféré de connexion (`PreferredLoginDevice`)** : Variable qualitative capturant le canal numérique majoritairement utilisé par le client (smartphone, ordinateur, etc.).
- **Classe de ville (`CityTier`)** : Niveau de développement de la zone géographique du client (tier 1, tier 2 ou tier 3), souvent corrélé au pouvoir d'achat et aux préférences d'achat.
- **Distance entre entrepôt et domicile (`WarehouseToHome`)** : Donnée quantitative susceptible d'influencer le délai de livraison, et donc la satisfaction client.
- **Mode de paiement préféré (`PreferredPaymentMode`)** : Variable catégorielle précisant le canal de paiement favori (carte bancaire, UPI, etc.), pouvant orienter des stratégies de personnalisation.
- **Genre (`Gender`)** : Attribut sociodémographique classique, servant aux études de segmentation.
- **Temps passé sur l'application (`HourSpendOnApp`)** : Nombre d'heures passées sur l'application mobile ou le site web, indicateur du niveau d'engagement numérique.
- **Nombre d'appareils enregistrés (`NumberOfDeviceRegistered`)** : Reflète la pluralité des points de contact d'un client avec la plateforme.
- **Catégorie préférée de commande (`PreferedOrderCat`)** : Type de produits achetés le plus fréquemment durant le mois précédent, utile pour les analyses comportementales.
- **Score de satisfaction (`SatisfactionScore`)** : Évaluation subjective du service par le client, généralement sur une échelle de 1 à 5. Variable à forte valeur explicative dans la modélisation du churn.
- **État matrimonial (`MaritalStatus`)** : Variable catégorielle pouvant influencer les décisions d'achat ou la fidélité, notamment dans les approches de marketing relationnel.
- **Nombre d'adresses enregistrées (`NumberOfAddress`)** : Donnée logistique utile pour analyser la mobilité et la complexité des préférences de livraison.
- **Réclamation (`Complain`)** : Indicateur binaire signalant si une plainte a été émise le mois précédent. Elle est souvent révélatrice d'un désengagement latent.
- **Évolution du montant des commandes (`OrderAmountHikeFromlastYear`)** : Pourcentage d'augmentation ou de diminution du montant commandé par rapport à l'année précédente. Sert à détecter les signaux faibles d'évolution du comportement d'achat.
- **Utilisation de coupons (`CouponUsed`)** : Fréquence d'usage des codes de réduction. Elle permet de cerner les clients sensibles aux incitations commerciales.
- **Volume de commandes (`OrderCount`)** : Nombre total de commandes effectuées le mois précédent. Mesure directe de l'activité transactionnelle.
- **Nombre de jours depuis la dernière commande (`DaySinceLastOrder`)** : Peut être interprété comme un indicateur de désengagement progressif.
- **Montant de cashback moyen (`CashbackAmount`)** : Moyenne mensuelle des remboursements ou remises crédités sur le compte du client. Paramètre potentiellement lié à la fidélité et à la rétention.

5.1.3 Typologie des variables

Sur le plan statistique, les variables se répartissent comme suit :

- **Variables numériques continues** : `Tenure`, `WarehouseToHome`, `HourSpendOnApp`, `OrderAmountHikeFromlastYear`, `CashbackAmount`, `DaySinceLastOrder`.
- **Variables discrètes (entiers)** : `NumberOfDeviceRegistered`, `NumberOfAddress`, `CouponUsed`, `OrderCount`, `SatisfactionScore`.
- **Variables catégorielles nominales** : `PreferredLoginDevice`, `PreferredPaymentMode`, `PreferedOrderCat`, `Gender`, `MaritalStatus`.
- **Variables binaires** : `Churn`, `Complain`.

5.1.4 Justification scientifique du choix du dataset

La pertinence du jeu de données retenu pour l'étude du churn s'appuie sur un ensemble d'éléments à la fois structurels, théoriques et opérationnels. Tout d'abord, la présence explicite d'une variable cible binaire (`Churn`), indiquant si un client a quitté la plateforme ou non, permet d'encadrer rigoureusement le problème comme une tâche de classification supervisée, conformément aux meilleures pratiques en analyse prédictive [1-5]. Cette approche est validée par *Lemon & Verhoef (2016)* dans leur revue des modèles de rétention client *(*Journal of Marketing*).*

Par ailleurs, les variables explicatives couvrent de manière exhaustive les facteurs traditionnellement associés au phénomène de désabonnement, tels que le niveau de satisfaction (`SatisfactionScore`), les réclamations récentes (`Complain`), la distance logistique entre l'entrepôt et le domicile (`WarehouseToHome`) [2-5], ainsi que des indicateurs comportementaux comme la récence (`DaySinceLastOrder`) et la fréquence des achats (`OrderCount`) [*Blattberg et al., 2008, Customer Equity]*. Ces dimensions sont reconnues dans la littérature comme des prédicteurs majeurs du churn.

Le jeu de données offre également une richesse comportementale appréciable : des informations sur le temps passé sur l'application (`HourSpendOnApp`), les préférences en matière de produits (`PreferedOrderCat`), ou encore les moyens de paiement utilisés (`PreferredPaymentMode`) permettent de capter des signaux faibles d'engagement ou de désintérêt [3-5],, souvent invisibles dans des données plus agrégées. Ces éléments facilitent l'implémentation de modèles à forte capacité explicative, comme les arbres de décision boostés ou les approches bayésiennes.

En outre, la nature temporelle de certaines variables telles que l'ancienneté (`Tenure`) ou l'inactivité (`DaySinceLastOrder`) rend ce jeu de données particulièrement propice à l'application de modèles de survie [4-5], qui permettent non seulement de prédire le churn, mais aussi d'en estimer le moment probable. Cette dimension temporelle s'avère précieuse pour coupler la prédiction à la valeur économique projetée du client (CLV), à partir de variables comme `OrderCount`, `CouponUsed` ou encore `CashbackAmount`.

Enfin, la structure tabulaire claire et propre du dataset, ainsi que l'hétérogénéité bien définie de ses variables, le rendent parfaitement exploitable dans des environnements analytiques modernes [5-5],. Il se prête tout autant à des visualisations interactives via Streamlit qu'à l'évaluation de la rentabilité prédictive à travers des métriques orientées business, telles que l'Expected Maximum Profit (EMP).

En résumé, ce jeu de données constitue un socle robuste, fidèle aux problématiques concrètes du e-commerce, et permet de mobiliser des approches statistiques et algorithmiques avancées, assurant ainsi la rigueur scientifique et la portée opérationnelle du projet.

5.2 Analyse exploratoire des données

5.2.1 Statistiques descriptives initiales

Afin d'obtenir une première vision quantitative du jeu de données, une analyse statistique descriptive a été effectuée sur l'ensemble des variables numériques. La commande `df.describe().transpose()` de la bibliothèque **Pandas** a permis de calculer, pour chaque variable, le nombre d'observations (`count`), la moyenne (`mean`), l'écart-type (`std`), les valeurs minimale et maximale (`min`, `max`), ainsi que les quartiles (`25%`, `50%` et `75%`). Le tableau ci-dessous présente les résultats obtenus :

	count	mean	std	min	25%	50%	75%	max
CustomerID	5630.0	52815.5	1625.39	50001.0	51408.25	52815.5	54222.75	55630.0
Churn	5630.0	0.17	0.37	0.0	0.0	0.0	0.0	1.0
Tenure	5366.0	10.19	8.56	0.0	2.0	9.0	16.0	61.0
CityTier	5630.0	1.65	0.92	1.0	1.0	1.0	3.0	3.0
WarehouseToHome	5379.0	15.64	8.53	5.0	9.0	14.0	20.0	127.0
HourSpendOnApp	5375.0	2.93	0.72	0.0	2.0	3.0	3.0	5.0
NumberOfDeviceRegistered	5630.0	3.69	1.02	1.0	3.0	4.0	4.0	6.0
SatisfactionScore	5630.0	3.07	1.38	1.0	2.0	3.0	4.0	5.0
NumberOfAddress	5630.0	4.21	2.58	1.0	2.0	3.0	6.0	22.0
Complain	5630.0	0.28	0.45	0.0	0.0	0.0	1.0	1.0
OrderAmountHikeFromlastYear	5365.0	15.71	3.68	11.0	13.0	15.0	18.0	26.0
CouponUsed	5374.0	1.75	1.89	0.0	1.0	1.0	2.0	16.0
OrderCount	5372.0	3.01	2.94	1.0	1.0	2.0	3.0	16.0
DaySinceLastOrder	5323.0	4.54	3.65	0.0	2.0	3.0	7.0	46.0
CashbackAmount	5630.0	177.22	49.21	0.0	145.77	163.28	196.39	324.99

En parallèle, une analyse des valeurs manquantes a été menée à l'aide des fonctions isnull() et sum() de **Pandas**.
Le jeu de données comporte un total de **1 856 valeurs manquantes**, soit **32,97 %** de l'ensemble des enregistrements. La répartition des valeurs manquantes par variable est présentée ci-dessous :

Variable	Valeurs manquantes
CustomerID	0
Churn	0
Tenure	264
PreferredLoginDevice	0
CityTier	0
WarehouseToHome	251
PreferredPaymentMode	0
Gender	0
HourSpendOnApp	255
NumberOfDeviceRegistered	0
PreferedOrderCat	0
SatisfactionScore	0
MaritalStatus	0
NumberOfAddress	0
Complain	0
OrderAmountHikeFromlastYear	265
CouponUsed	256
OrderCount	258
DaySinceLastOrder	307
CashbackAmount	0

L'examen de ces résultats montre que certaines variables présentent un volume significatif de données manquantes, notamment `Tenure`, `WarehouseToHome`, `HourSpendOnApp`, `OrderAmountHikeFromlastYear`, `CouponUsed`, `OrderCount` et `DaySinceLastOrder`. Ces variables nécessitent un traitement approprié lors de la phase de **prétraitement des données** afin d'éviter toute perte d'information ou introduction de biais dans les modèles d'apprentissage automatique.

5.2.2 Pipeline de traitement des données

Introduction

Le traitement préalable des données constitue une phase critique dans la mise en œuvre d'un modèle prédictif robuste et interprétable, en particulier dans le contexte de la fidélisation client. L'ensemble des étapes de ce pipeline a été construit pour maximiser la qualité de l'information, limiter les biais, et garantir la reproductibilité des résultats tout en respectant les exigences d'un cadre industriel e-commerce.

5.2.2.1 Prétraitement global des données

Gestion des doublons et des valeurs manquantes

Une première vérification a été effectuée afin de détecter d'éventuels doublons dans le jeu de données. L'identifiant unique `CustomerID` a été utilisé comme référence, et aucun doublon n'a été détecté, garantissant ainsi l'unicité des enregistrements clients.

En ce qui concerne les valeurs manquantes (voir section *Statistiques descriptives initiales et valeurs manquantes*), plusieurs variables présentaient des données absentes. Deux stratégies d'imputation ont été retenues, en fonction de la nature et de la distribution des variables :

- **Pour la variable `DaySinceLastOrder`** : les valeurs manquantes ont été remplacées par la valeur maximale observée dans cette variable. Ce choix se justifie par le fait que l'absence d'information sur le nombre de jours depuis la dernière commande peut raisonnablement indiquer que le client n'a pas passé de commande depuis une très longue période. Ainsi, l'imputation par la valeur maximale permet de conserver cette interprétation sans introduire de biais vers des valeurs plus faibles.
- **Pour les autres variables présentant des valeurs manquantes** (`Tenure`, `WarehouseToHome`, `HourSpendOnApp`, `OrderAmountHikeFromlastYear`, `CouponUsed`, `OrderCount`), l'imputation a été réalisée à l'aide de la **médiane**. Ce choix est motivé par les résultats des statistiques descriptives : les distributions de ces variables présentent des valeurs extrêmes susceptibles de biaiser la moyenne. La médiane, moins sensible aux outliers, constitue donc un estimateur plus robuste de tendance centrale dans ce contexte [6-5]. Cette approche permet de conserver la cohérence structurelle du jeu de données sans altérer la variance globale.

Cette procédure garantit que l'ensemble des valeurs manquantes est traité de manière cohérente, tout en minimisant les risques d'introduire un biais important dans les étapes ultérieures d'analyse et de modélisation.

Détection et traitement des valeurs aberrantes (Outliers)

Dans le cadre de l'analyse exploratoire des données, une attention particulière a été portée à la détection des valeurs aberrantes (outliers) pouvant potentiellement perturber la qualité des modèles prédictifs. Toutefois, compte tenu de la nature et de la distribution des variables, l'étude s'est concentrée uniquement sur trois variables numériques clés :

- `NumberOfAddress`
- `Tenure`
- `WarehouseToHome`

Cette sélection s'explique par le fait que certaines variables numériques du dataset présentent une distribution discontinue ou des valeurs extrêmes considérées comme acceptables dans le contexte métier. Par exemple, des variables comme `OrderCount` ou `CouponUsed` peuvent naturellement comporter des valeurs élevées, sans qu'elles soient nécessairement aberrantes.

Méthode de détection des outliers

La méthode retenue repose sur la règle classique de l'**intervalle interquartile (IQR)** :

$IQR = Q_3 - Q_1$

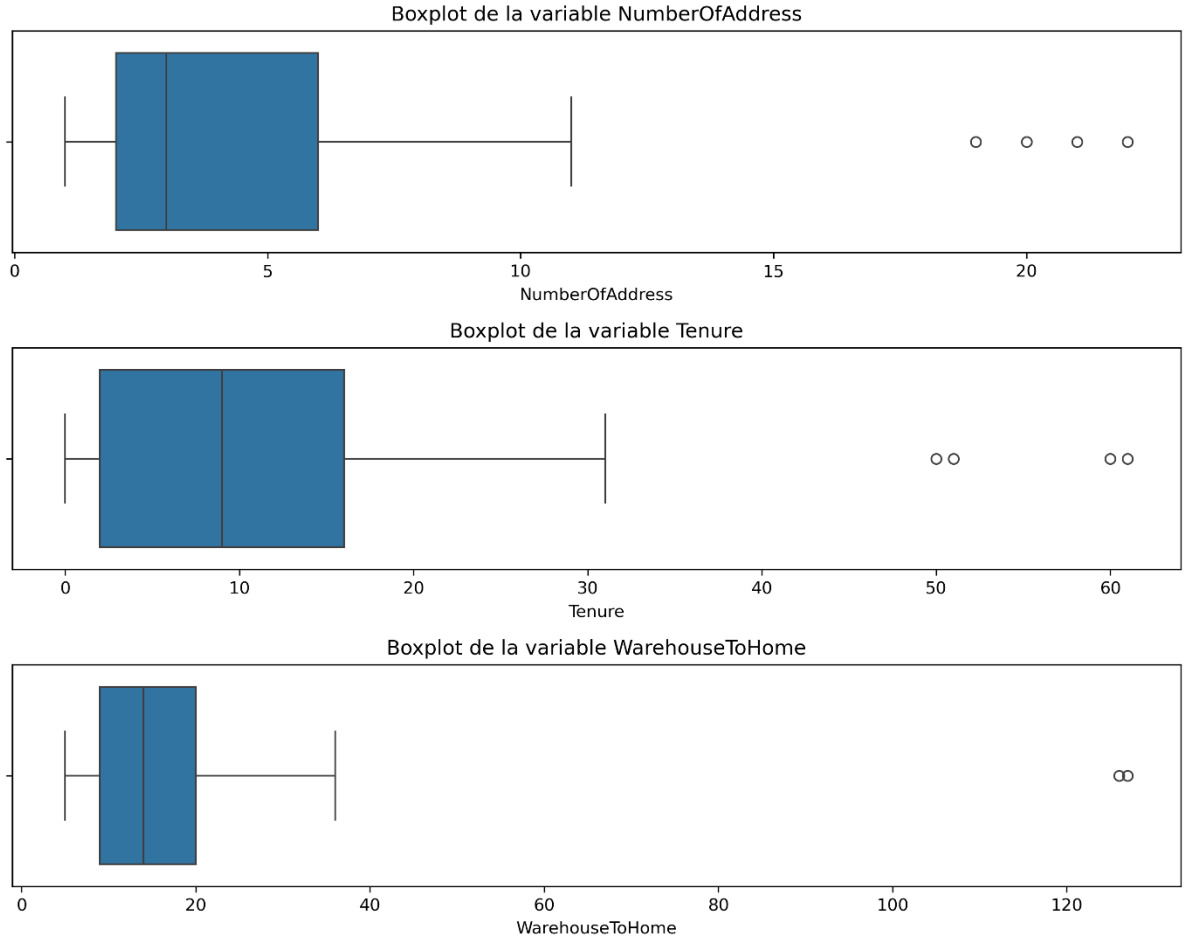
où Q_1 et Q_3 sont respectivement le premier et le troisième quartile.
Les observations situées en dehors de l'intervalle $[Q_1 - 1.5 \times IQR ; Q_3 + 1.5 \times IQR]$ sont considérées comme des outliers.

Résultats de la détection

L’analyse a conduit aux résultats suivants pour les variables sélectionnées :

Variable	Nombre d’outliers détectés
NumberOfAddress	4
Tenure	4
WarehouseToHome	2

Les boxplots ci-dessous illustrent graphiquement la répartition des données et la présence des outliers détectés sur ces variables :



Traitement des outliers

Pour limiter l’impact des valeurs aberrantes sur les modèles, les outliers identifiés ont été remplacés par la médiane de la variable correspondante, méthode robuste et simple qui préserve la structure générale des données. Cette approche a permis de ramener le nombre d’outliers détectés à zéro pour ces variables après traitement.

Néanmoins, il est important de noter que certaines variables, bien qu’ayant présenté des outliers (notamment `OrderCount` avec 703 outliers ou `CashbackAmount` avec 438 outliers), n’ont pas été modifiées. Cette décision repose sur la compréhension métier et la nature de ces variables, où des valeurs extrêmes peuvent être justifiées par des comportements clients légitimes et ne

constituent pas nécessairement des erreurs ou anomalies. Un traitement plus spécifique pourra être envisagé dans une phase ultérieure selon les objectifs analytiques.

Encodage des variables catégorielles

Dans le cadre de l'apprentissage automatique pour la prédiction de l'attrition client, de nombreux algorithmes requièrent que les données en entrée soient représentées sous forme numérique afin de pouvoir effectuer des opérations arithmétiques ou statistiques. Les variables catégorielles, par nature qualitatives, ne peuvent donc pas être traitées directement par ces algorithmes. L'encodage numérique de ces variables constitue ainsi une étape essentielle de la phase de prétraitement des données, visant à rendre l'information exploitable tout en préservant la correspondance exacte entre chaque modalité et sa représentation codée.

Dans ce contexte, deux méthodes d'encodage sont mobilisées dans cette étude : **le Label Encoding** et le **One-Hot Encoding**, chacune répondant à des besoins spécifiques en fonction du nombre et de la nature des modalités.

Rappel théorique

Label Encoding

Le **Label Encoding** est une technique d'encodage ordinal consistant à attribuer un entier unique à chaque modalité distincte d'une variable catégorielle. Soit

$$\mathcal{C} = \{c_1, c_2, \dots, c_k\}$$

L'ensemble des k modalités distinctes d'une variable catégorielle. Le Label Encoding définit une application bijective :

$$f: \mathcal{C} \longrightarrow \mathbb{N}_k = \{0,1, \dots, k-1\}, \qquad f(c_i) = j$$

où j est l'indice associé à la modalité c_i après identification et tri (souvent lexicographique) des modalités. L'opération inverse, permettant de retrouver la valeur initiale, est donnée par :

$$f^{-1}: \mathbb{N}_k \longrightarrow \mathcal{C}\ , f^{-1}(j) = c_i$$

Ce mécanisme permet de convertir rapidement des variables qualitatives en un format numérique compact, adapté notamment aux algorithmes ne tenant pas compte d'une éventuelle distance entre les codes (p. ex. : arbres de décision).

Limites : l'encodage ainsi produit induit artificiellement une structure d'ordre ($0<1<\cdots<k-1$) entre les modalités, ce qui peut engendrer un biais dans les modèles sensibles aux distances et aux relations ordinales, tels que la régression linéaire, les SVM ou le k-NN.

Afin de lever cette contrainte, en particulier pour les variables catégorielles présentant plus de trois modalités, le **One-Hot Encoding** est utilisé. Cette méthode transforme chaque modalité en une variable binaire distincte, évitant ainsi l'introduction d'un ordre artificiel entre les catégories.

One-Hot Encoding.

Le **One-Hot Encoding** est une technique d'encodage binaire qui permet de représenter chaque modalité d'une variable catégorielle par un vecteur unitaire, éliminant ainsi tout risque d'introduction d'un ordre artificiel entre les catégories.

Formellement, soit $\mathcal{C} = \{c_1, c_2, \dots, c_k\}$ l'ensemble des modalités distinctes d'une variable catégorielle. Le One-Hot Encoding définit une application :

$$g: \mathcal{C} \longrightarrow \{0,1\}^k$$

telle que :

$$g(c_i) = (0, \dots, 0, 1, 0, \dots, 0)$$

où le 1 est placé à la i-ème position et toutes les autres composantes sont nulles.

Par exemple, pour C={Rouge, Vert, Bleu} :

g(Rouge)=(1,0,0), g(Vert)=(0,1,0), g(Bleu)=(0,0,1)

Cette représentation permet aux algorithmes de traiter les variables catégorielles **sans leur attribuer de relation d'ordre implicite**, tout en maintenant une information parfaitement fidèle sur la modalité initiale.

Avantages : elle supprime toute ambiguïté d'interprétation due à l'encodage numérique ordinal et s'adapte particulièrement bien aux modèles linéaires et aux algorithmes utilisant des distances euclidiennes.

Inconvénients : le principal inconvénient réside dans l'augmentation dimensionnelle : une variable avec k modalités est transformée en k nouvelles variables binaires, ce qui peut accroître le coût computationnel et nécessiter des ressources mémoire supplémentaires, en particulier lorsque k est grand.

L'objectif de l'encodage one-hot est de représenter les modalités sans introduire d'ordonnancement artificiel ou de biais numérique. Il permet ainsi aux algorithmes de classification ou de régression, notamment ceux basés sur des distances (KNN, SVM) ou sur la construction d'hyperplans (régression logistique, réseaux de neurones), d'interpréter correctement les caractéristiques non ordinales [15-4].

Application.

Les variables catégorielles ont été traitées selon leur nature. La variable `Gender`, binaires, ont été encodées à l'aide de `LabelEncoder`, ce qui est suffisant dès lors qu'il n'existe que deux modalités sans ordre implicite [8-5].

En revanche, les variables à plusieurs modalités non ordinales (`PreferredLoginDevice`, `PreferredPaymentMode`, `PreferedOrderCat` et `MaritalStatus`) ont été traitées par **encodage One-Hot**, avec suppression de la première modalité (`drop_first=True`) pour éviter la redondance linéaire (piège de la dummy variable). Le choix de ne pas appliquer `LabelEncoder` à ces variables est motivé par l'absence d'ordre naturel entre les modalités : un encodage numérique arbitraire induirait une hiérarchisation artificielle [9-5], susceptible de biaiser les modèles.

Enfin, les colonnes de type booléen (issues d'un prétraitement antérieur ou de transformations) ont été converties en type entier (`int`) afin d'assurer leur compatibilité avec les modèles d'apprentissage automatique.

Standarisation des données

La standardisation constitue une étape cruciale du prétraitement des données dans les projets d'apprentissage automatique, notamment lorsque les variables présentent des échelles ou des unités hétérogènes. En effet, de nombreux algorithmes, tels que les méthodes basées sur la distance (k-NN, SVM), les modèles linéaires ou encore les réseaux de neurones, sont sensibles à la variance et à la distribution des variables en entrée [7-5]. Des variables non normalisées peuvent dominer le processus d'apprentissage, biaisant ainsi les résultats et dégradant la performance des modèles.

La standardisation vise à transformer les variables numériques afin de leur conférer une distribution centrée réduite, c'est-à-dire une moyenne nulle et un écart-type unitaire.

Mathématiquement, pour une variable $X=(x_1,x_2,\ldots,x_n)$, la variable standardisée $Z=(z_1,z_2,\ldots,z_n)$ est calculée par :

$$z_i = \frac{x_i - \mu_X}{\sigma_X}$$

où μ_X est la moyenne empirique de X et σ_X son écart-type empirique.

Cette transformation assure que chaque variable contribue de manière équitable dans les calculs impliquant des distances ou des coefficients, facilitant ainsi la convergence des algorithmes d'apprentissage et améliorant leur robustesse.

Remarque importante : la standardisation est appliquée uniquement sur les variables numériques continues ou discrètes non binaires. Les variables binaires (par exemple, `Gender`, `Complain`) et les variables ordinales à faible nombre de modalités (comme `CityTier`) sont exclues, afin de préserver la signification intrinsèque et l'échelle de ces données.

Dans cette étude, la standardisation a été appliquée aux variables suivantes :

- `Tenure`
- `WarehouseToHome`
- `HourSpendOnApp`
- `NumberOfDeviceRegistered`
- `NumberOfAddress`
- `OrderAmountHikeFromlastYear`
- `OrderCount`
- `DaySinceLastOrder`
- `CashbackAmount`

L'implémentation a été réalisée à l'aide de la classe `StandardScaler` du module `sklearn.preprocessing`.

5.3 Sélection des variables

Introduction

L'ensemble de données sur le churn client contenait de nombreuses caractéristiques clients surtout après l'encodage par One-Hot Encoding, et toutes les variables n'étaient pas propices à la performance prédictive. Un excès de redondances et de variables non pertinentes dans le jeu de données peut freiner la capacité prédictive du modèle [Ref- Verbeke, W.; Dejaeger, K.; Martens, D.; Hur, J.; Baesens, B. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. Eur. J. Oper. Res. 2012, 218, 211–229]. C'est pourquoi cette étape a été effectuée.
La sélection des variables est une étape cruciale du prétraitement des données en apprentissage automatique, notamment pour la prédiction du churn. Elle permet de réduire la dimensionnalité, d'améliorer la performance des modèles, de diminuer le risque de surapprentissage (*overfitting*) et de faciliter l'interprétation des résultats. Parmi les méthodes de sélection existantes, **Random Forest** a été choisi pour cette étude.

La **forêt aléatoire (Random Forest)** est un algorithme efficace de sélection de variables, présentant une forte précision de classification, une bonne robustesse face au bruit et aux valeurs aberrantes, ainsi qu'une grande capacité de généralisation [Ref Breiman, L. Random forests. Mach. Learn. 2001, 45, 5–32.]

s'impose comme un outil robuste grâce à sa capacité à mesurer l'importance des variables de manière intrinsèque, sans nécessiter de transformation préalable du jeu de données.

Rappel théorique du mécanisme

Un **Random Forest** est un ensemble d'arbres de décision $h_1, h_2, \ldots, h_T$, construits à partir de sous-échantillons aléatoires des données et de sous-ensembles aléatoires de variables. L'importance d'une variable X_j est évaluée en mesurant la réduction moyenne de l'impureté

(par exemple, l'indice de Gini [10-5]) qu'elle apporte lors des séparations dans les arbres. Mathématiquement, pour une variable X_j , l'importance est définie par :

$$Importance(X_j) = \frac{1}{T}\sum_{1}^{T}\sum_{n \in N_{j,t}} p(n) \cdot \triangle\, i(n)$$

où :

- Test le nombre d'arbres,
- $N_{j,\,t}$ est l'ensemble des nœuds de l'arbre t où X_j est utilisée,
- p(n) est la proportion d'échantillons arrivant au nœud n,
- Δi(n) est la réduction de l'impureté produite par la division au nœud n.

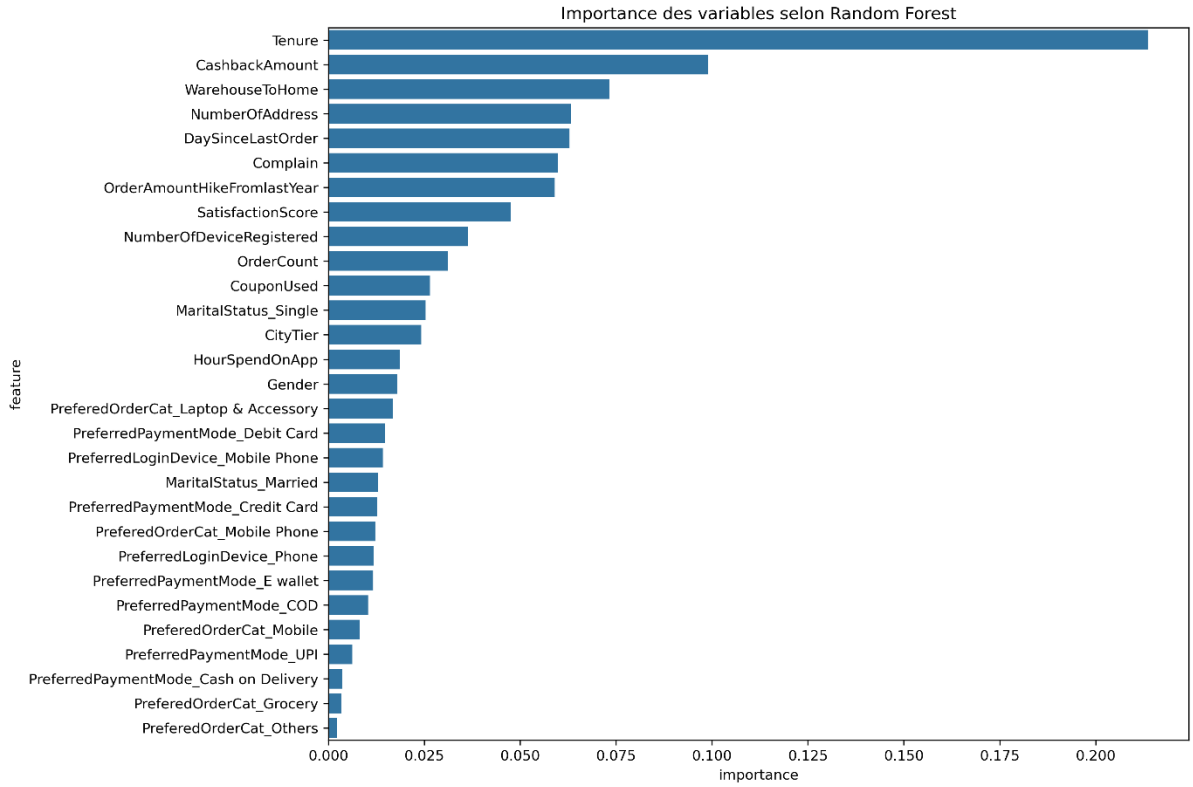
Cette mesure est normalisée afin que la somme des importances sur toutes les variables soit égale à 1.

Résultats et interprétation

Le jeu de données initial comportait **19 variables explicatives**. Après application des techniques d'encodage présentées précédemment, le nombre de colonnes est passé à **30**. L'analyse par **Random Forest** a produit le classement suivant des variables par ordre d'importance (Top 10 affiché ci-dessous) :

Rang	Variable	Importance
1	Tenure	0.2136
2	CashbackAmount	0.0989
3	WarehouseToHome	0.0732
4	NumberOfAddress	0.0633
5	DaySinceLastOrder	0.0629
6	Complain	0.0598
7	OrderAmountHikeFromlastYear	0.0590
8	SatisfactionScore	0.0475
9	NumberOfDeviceRegistered	0.0364
10	OrderCount	0.0311

Et voici le classement des variables en images :



L'analyse montre que **Tenure** est de loin la variable la plus discriminante pour la prédiction de la cible, suivie par **CashbackAmount** et **WarehouseToHome**. À l'inverse, certaines variables telles que *PreferedOrderCat_Others* (0.0023), *PreferedOrderCat_Grocery* (0.0034) ou *PreferredPaymentMode_Cash on Delivery* (0.0036) présentent des importances négligeables.

Conclusion opérationnelle

Sur la base de ces résultats, il est possible de réduire la dimensionnalité du jeu de données en supprimant les variables les moins contributives. Ainsi, les trois dernières du classement, à savoir:

- **PreferedOrderCat_Others** (0.0023)
- **PreferedOrderCat_Grocery** (0.0034)
- **PreferredPaymentMode_Cash on Delivery** (0.0036)

pourraient être retirées du modèle, ce qui permettrait de simplifier l'espace des caractéristiques tout en préservant la performance prédictive.

Les variables les plus influentes sont `Tenure`, `SatisfactionScore`, `CashbackAmount`, `Complain` et `OrderCount`, ce qui corrobore les résultats de la littérature sur les déterminants comportementaux et transactionnels du désabonnement client.

5.4 Implémentation des modèles prédictifs de classification

5.4. 1 Segmentation avant la classification

Avant l'entraînement des modèles de classification pour la prédiction du churn, une étape de segmentation des clients a été réalisée à l'aide de l'algorithme **K-Means**, permettant de créer une nouvelle variable **Segment**. Cette étape, largement adoptée dans la littérature récente, présente plusieurs avantages tant méthodologiques que pratiques.

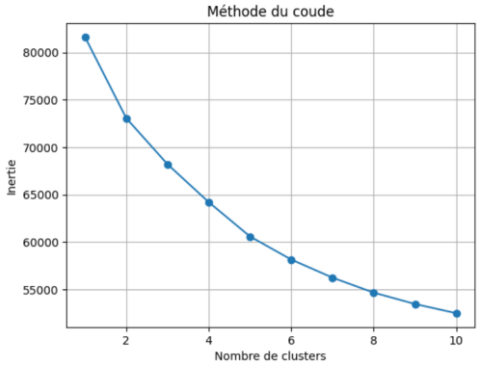
D'après plusieurs études, la segmentation via K-Means a été appliquée avec succès dans différents secteurs tels que le e-commerce [36-chap5-4], le retail [37-chap5-4], la finance [*40-chap5-4*] et les télécommunications [41-chap5-4]. Sa popularité s'explique par sa simplicité, sa rapidité de calcul et sa capacité à détecter des structures naturelles au sein de données multidimensionnelles.

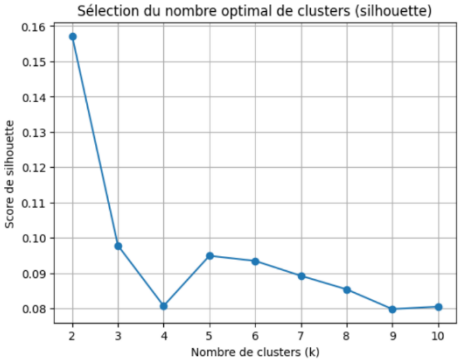
Une autre justification repose sur la nature des churns. Certains clients peuvent cesser leurs achats pour des raisons indépendantes de l'entreprise, comme la perte d'emploi, un déménagement ou le décès. La segmentation permet alors de détecter des profils présentant des comportements atypiques ou des churns involontaires, ce qui contribue à une meilleure compréhension et prise en compte des différents types de churn.

Par ailleurs, comme le souligne l'article 2 [B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM Xiancheng Xiahou * and Yoshio Harada], intégrer une étape de segmentation avant l'application des modèles de classification, tels que la régression logistique ou le SVM, **améliore la performance prédictive**. En regroupant les clients selon des comportements homogènes, la variance intra-cluster est réduite, ce qui facilite l'apprentissage des relations entre les caractéristiques et le churn.

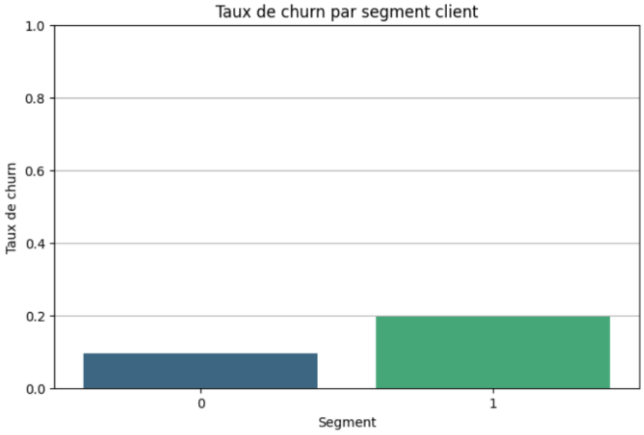
La segmentation offre également un avantage opérationnel : elle permet de concevoir des stratégies marketing ciblées et adaptées à chaque profil, plutôt que de traiter l'ensemble des clients comme un groupe homogène.

Pour déterminer le **nombre optimal de clusters**, nous avons combiné la **méthode du coude** et le **coefficient de silhouette** (--------FIGURE X----)**.** Ces analyses ont conduit à retenir **deux clusters**, et la colonne **Segment** a été créée avec les valeurs 0 et 1, représentant l'appartenance de chaque client à l'un des deux clusters.





Le résultats :



5.4 .2 Partitionnement des données et gestion des classes déséquilibrées

La séparation du jeu de données en échantillons d'entraînement et de test a été réalisée dans une proportion **80/20**, en appliquant l'option `stratify=y`. Cette stratification garantit que la répartition de la classe cible (`Churn`) est conservée dans les deux sous-ensembles, une approche validée pour préserver la distribution des classes, afin d'évaluer la performance réelle du modèle sur une distribution représentative.

Gestion des classes déséquilibrées

Problèmes de classes déséquilibrées en churn

Dans les problématiques de **prédiction du churn**, on observe fréquemment une **forte asymétrie dans la répartition des classes** : les clients ayant quitté le service représentent souvent une **minorité** comparée à ceux qui restent actifs. Ce déséquilibre peut conduire les modèles d'apprentissage supervisé à être **biaisés en faveur de la classe majoritaire**, au détriment de la détection efficace des clients à risque.

En effet, sans traitement spécifique, un modèle peut obtenir une **exactitude globale élevée**, tout en négligeant quasiment totalement la classe minoritaire. Cette situation est particulièrement problématique dans le cas du churn, où l'objectif principal est précisément de **détecter précocement les clients susceptibles de se désengager**. Il est donc indispensable d'adopter

des stratégies permettant de **rééquilibrer la distribution des classes** pour améliorer la sensibilité et la pertinence du modèle vis-à-vis des clients churners.

Méthode SMOTE

Dans le contexte d'un déséquilibre marqué entre classes, comme c'est souvent le cas pour la prédiction du churn, la méthode **SMOTE** (*Synthetic Minority Over-sampling Technique*) constitue une solution couramment adoptée pour améliorer la performance des modèles d'apprentissage automatique. Développée par Chawla et al. (2002), cette technique vise à **augmenter la représentativité de la classe minoritaire** en générant de manière artificielle de nouveaux exemples.

Le principe de SMOTE repose sur l'**interpolation linéaire entre une observation de la classe minoritaire et l'un de ses k plus proches voisins** (également de la classe minoritaire). Contrairement à un simple sur-échantillonnage par duplication, cette approche introduit une **variabilité contrôlée** dans les données, ce qui permet au modèle d'apprentissage d'**appréhender plus finement la frontière de décision**.

Bien que SMOTE soit efficace pour **réduire le biais de classification en faveur de la classe majoritaire**, elle présente certaines **limites importantes** :

- Elle peut générer des exemples synthétiques situés dans des zones peu représentatives ou ambiguës, en particulier lorsque les classes sont très imbriquées.
- Elle **ne prend pas en compte la classe majoritaire** lors de la création de nouveaux points, ce qui peut conduire à la création de données synthétiques qui se chevauchent avec l'autre classe.
- Elle est sensible aux **bruits et anomalies** présents dans la classe minoritaire, qui peuvent être amplifiés par le sur-échantillonnage.

Malgré ces limites, SMOTE reste une technique robuste et accessible, largement utilisée dans les applications industrielles et académiques où le déséquilibre de classes compromet la performance des modèles prédictifs.

Application

Le jeu de données présente un déséquilibre manifeste entre les classes : seulement **16.84 % des clients** ont churné, soit **948 individus sur 5 630**. Un tel déséquilibre peut conduire les modèles supervisés à ignorer la classe minoritaire au profit de la majorité [11-5], dégradant ainsi les performances en détection de churn.

Pour pallier cet effet, nous avons recours à **SMOTE (Synthetic Minority Oversampling Technique),** notamment pour les données d'entrainement. Contrairement à un suréchantillonnage aléatoire qui duplique les observations existantes, SMOTE génère de **nouvelles instances synthétiques** à partir des voisins proches de la classe minoritaire dans

l'espace des variables [12-5]. Cette méthode enrichit la variété structurelle des churners et améliore la capacité généralisante des modèles [12-5], sans introduire de surapprentissage.

5.4.3 Paramètrage et entrainement des modèles

Afin de prédire le churn des clients, plusieurs modèles de classification ont été implémentés et comparés. L'ensemble des expériences a été réalisé sur les mêmes données d'entraînement, enrichies de la variable de segmentation **Segment** issue du K-Means, à l'exception de certains cas où cette variable a volontairement été exclue pour évaluer son impact.

Prétraitement des données

Dans un premier temps, trois variables considérées comme les moins influentes par l'algorithme de sélection de caractéristiques via Random Forest ont été supprimées :

- *PreferedOrderCat_Others*
- *PreferedOrderCat_Grocery*
- *PreferredPaymentMode_Cash on Delivery*

De plus, l'identifiant du client (*CustomerID*), n'ayant pas de pertinence prédictive, a également été retiré.

Les données ont ensuite été séparées en ensembles d'apprentissage et de test selon une proportion 80/20, avec une stratification sur la variable cible afin de conserver la distribution initiale des classes. Pour pallier le problème de déséquilibre entre churners et non-churners, la technique **SMOTE (Synthetic Minority Oversampling Technique)** a été appliquée sur l'ensemble d'entraînement, permettant de générer artificiellement de nouveaux exemples de la classe minoritaire et d'améliorer ainsi la robustesse des modèles.

Modèles implémentés

Cinq modèles principaux ont été entraînés :

- **Régression logistique** : modèle de base, souvent utilisé pour la prédiction binaire du churn, afin de disposer d'un point de comparaison.
- **SVM (Support Vector Machine)** : pour sa capacité à bien séparer les classes, notamment dans des espaces de grande dimension.
- **XGBoost** : algorithme d'ensemble basé sur le boosting, réputé pour sa performance dans les compétitions de machine learning. Deux expériences ont été menées :
    1. **Avec et sans suppression des trois variables jugées peu influentes** afin d'évaluer l'impact de la réduction de dimension sur les performances.
    2. **Avec et sans la variable Segment** issue de la segmentation K-Means, pour tester l'apport de cette étape de clustering préalable.
- **Réseaux de neurones artificiels (ANN)** : deux architectures distinctes ont été testées, afin d'évaluer l'influence de la complexité du modèle et de ses hyperparamètres :
    - **ANN 1** : une architecture simple composée de deux couches cachées (64 et 32 neurones avec fonctions d'activation ReLU), des couches de régularisation par *dropout*, et une sortie sigmoïde pour la classification binaire. L'optimiseur *Adam* et la fonction de perte *binary cross-entropy* ont été utilisés, avec un arrêt anticipé (*early stopping*) basé sur la perte de validation.

- **ANN 2** : une architecture plus profonde et régularisée, avec trois couches cachées (128, 64, 32 neurones), des normalisations par batch (*BatchNormalization*) et des taux de dropout plus variés (0.5, 0.3 et 0.2). L'optimiseur Adam a été paramétré avec un **learning rate initial de 0.001**, et l'entraînement a été supervisé par des callbacks avancés (*early stopping* sur l'AUC, réduction adaptative du taux d'apprentissage). Les métriques suivies incluaient la précision, le rappel, l'AUC et l'accuracy.

Comparaisons effectuées

Afin d'analyser l'apport des différentes étapes méthodologiques, plusieurs comparaisons ont été réalisées :

1. **XGBoost avec et sans suppression des variables peu influentes** permet de mesurer l'intérêt de la sélection de caractéristiques par Random Forest.
2. **XGBoost avec et sans la variable Segment** permet d'évaluer la contribution de la segmentation préalable via K-Means à l'amélioration de la performance du modèle.
3. **ANN 1 vs ANN 2** : permet de tester l'effet de la profondeur et de la régularisation sur la capacité prédictive des réseaux de neurones.

Ces expériences ont permis d'obtenir une vision comparative claire entre des modèles linéaires, non linéaires et neuronaux, tout en tenant compte des apports de la segmentation et de la sélection de variables.

5.4.4 Comparaison des modèles

Métriques d'évaluation

Pour évaluer et comparer les différents modèles prédictifs de churn, nous avons retenu plusieurs métriques complémentaires :

- **F1-score** : mesure l'équilibre entre précision (*precision*) et rappel (*recall*). Cette métrique est essentielle dans le cas du churn, car elle reflète la capacité du modèle à identifier correctement les churners sans générer trop de faux positifs.
- **Précision** : indique la proportion de churns prédits correctement parmi l'ensemble des clients identifiés comme churners. Dans une logique de **marketing ciblé**, cette métrique est cruciale, car une précision élevée signifie que les campagnes promotionnelles seront dirigées vers de véritables churners et non vers des clients fidèles, réduisant ainsi les coûts.
- **AUC (Area Under the ROC Curve)** : mesure la capacité du modèle à distinguer correctement les classes. Une AUC proche de 1 indique un excellent pouvoir discriminant.
- **Matrice de confusion** : permet une visualisation directe des faux positifs et des faux négatifs, donnant une interprétation plus opérationnelle des résultats.

Résultats expérimentaux

Le tableau suivant synthétise les performances obtenues :

Modèle	Précision (classe 1)	F1-score (classe 1)	AUC (ROC)	Accuracy	Commentaire
Régression Logistique	0.44	0.55	0.85	0.80	Faible capacité à détecter les churners, beaucoup de faux positifs.
SVM	0.65	0.72	0.93	0.89	Bon compromis précision/rappel, meilleure séparation des classes que la régression logistique.
ANN 1	0.82	0.84	0.974	0.95	Réseau simple, bons résultats avec peu d'overfitting.
ANN 2	0.84	0.85	0.98	0.95	Architecture plus profonde, performances très stables et robustes.
XGBoost (avec segment + suppr colonnes)	0.98	0.96	≈0.999	0.99	Meilleur modèle, excellents résultats globaux.
XGBoost (sans segment, suppr colonnes)	0.96	0.95	0.9983	0.98	Performances élevées, mais légèrement inférieures à la version avec segment.
XGBoost (avec segment, sans suppr colonnes)	0.97	0.95	0.9987	0.98	Performances solides, proche du modèle final mais avec plus de variables.

Discussion des résultats

- **Régression Logistique** : ce modèle de base a montré ses limites avec une précision très faible (0.44) pour la classe *churn*, rendant son utilisation risquée dans un cadre opérationnel.
- **SVM** : amélioration significative par rapport à la régression logistique (F1 = 0.72), mais reste en retrait face aux réseaux de neurones et à XGBoost.
- **Réseaux de Neurones (ANN 1 et ANN 2)** : les deux architectures ont montré des performances très solides, avec des F1-scores supérieurs à 0.84. ANN 2, grâce à une architecture plus riche et des techniques de régularisation (BatchNorm, Dropout, callbacks avancés), s'est révélé légèrement meilleur et plus robuste.

  Toutefois, il est important de noter que les ANN pourraient encore être améliorés via un ajustement fin des hyperparamètres. Néanmoins, leur **temps d'exécution long** et la **taille relativement limitée du jeu de données** (moins de 6000 échantillons) constituent des freins importants, surtout face à l'efficacité et à la rapidité d'entraînement de XGBoost.

- **XGBoost** : ce modèle surpasse largement tous les autres modèles, atteignant un F1-score proche de 0.96 et une précision de 0.98. Il combine un excellent pouvoir discriminant (AUC ≈ 0.999) et une robustesse remarquable. De plus, l'ajout de la variable **Segment** améliore la performance par rapport à la version sans segmentation (F1 = 0.96 contre 0.95). Cela confirme l'hypothèse que la segmentation préalable aide à mieux capter les comportements différenciés des clients.
- **Impact de la suppression des colonnes** : la suppression des variables peu influentes identifiées par Random Forest n'a pas détérioré les résultats ; au contraire, elle a permis de simplifier le modèle tout en maintenant un très haut niveau de performance.

Choix final

Au regard de l'ensemble des résultats, **XGBoost avec la variable** *Segment* **et suppression des trois colonnes peu influentes** a été retenu comme **modèle final**.
Ce choix repose sur :

- une **précision très élevée (0.98)**, essentielle pour cibler efficacement les véritables churners,
- un **F1-score de 0.96**, garantissant un bon équilibre entre détection des churners et limitation des faux positifs,
- une **AUC ≈ 0.999**, confirmant l'excellent pouvoir discriminant du modèle,
- et une robustesse démontrée face à la variabilité des données d'entraînement.

En conclusion, ce modèle constitue la solution la plus adaptée pour la prédiction du churn dans le cadre de ce projet. Il peut être intégré directement dans une démarche de fidélisation client, avec un fort impact opérationnel grâce à sa capacité à réduire les coûts des campagnes marketing tout en améliorant leur efficacité.

5.5 Analyse de Survie

Introduction

Jusqu'à présent, les modèles de classification et de prédiction comme le SVM, le réseau de neurones ou encore l'algorithme XGBoost ont permis d'identifier les clients susceptibles de quitter l'entreprise. Cependant, ces approches restent limitées dans leur capacité à répondre à une question essentielle pour la stratégie de fidélisation : **« Quand » un client risque-t-il de partir ?** En effet, XGBoost et les méthodes de classification supervisée se focalisent sur la probabilité de churn mais n'apportent aucune information temporelle sur la durée de rétention. C'est dans ce contexte que l'**analyse de survie** trouve toute sa pertinence, en permettant d'estimer la probabilité de rester client en fonction du temps, tout en intégrant la dynamique temporelle du comportement de départ.

L'objectif de cette section est donc de compléter les résultats prédictifs obtenus précédemment par une modélisation temporelle de la durée de vie client, afin d'apporter une vision plus fine et plus opérationnelle pour l'entreprise.

Plusieurs approches sont disponibles pour cela :

- des méthodes **non paramétriques** comme Kaplan-Meier pour obtenir une vue empirique de la survie,
- des modèles **semi-paramétriques** comme le modèle de Cox à risques proportionnels, largement utilisé pour sa flexibilité, permettant de mesurer l'effet relatif de plusieurs covariables sur le risque instantané de churn,

- des modèles **paramétriques** comme le modèle de régression de Weibull, qui fournit une estimation analytique du temps de survie et une meilleure interprétabilité des coefficients.

Dans ce travail, nous avons exploré ces trois familles de modèles afin d'évaluer leurs performances respectives sur le jeu de données client, tant en termes de capacité prédictive que d'adéquation aux hypothèses sous-jacentes.

Pour comparer les performances, nous avons utilisé des métriques standard de la littérature en survie :

- **le C-index** (concordance index), mesurant la capacité du modèle à correctement ordonner les durées de survie ;
- **le Brier Score** à différents horizons de temps (3, 9, 15 mois), mesurant la précision des probabilités de survie prédite ;
- **l'Integrated Brier Score (IBS)**, représentant la moyenne pondérée du Brier Score sur toute la durée de suivi.

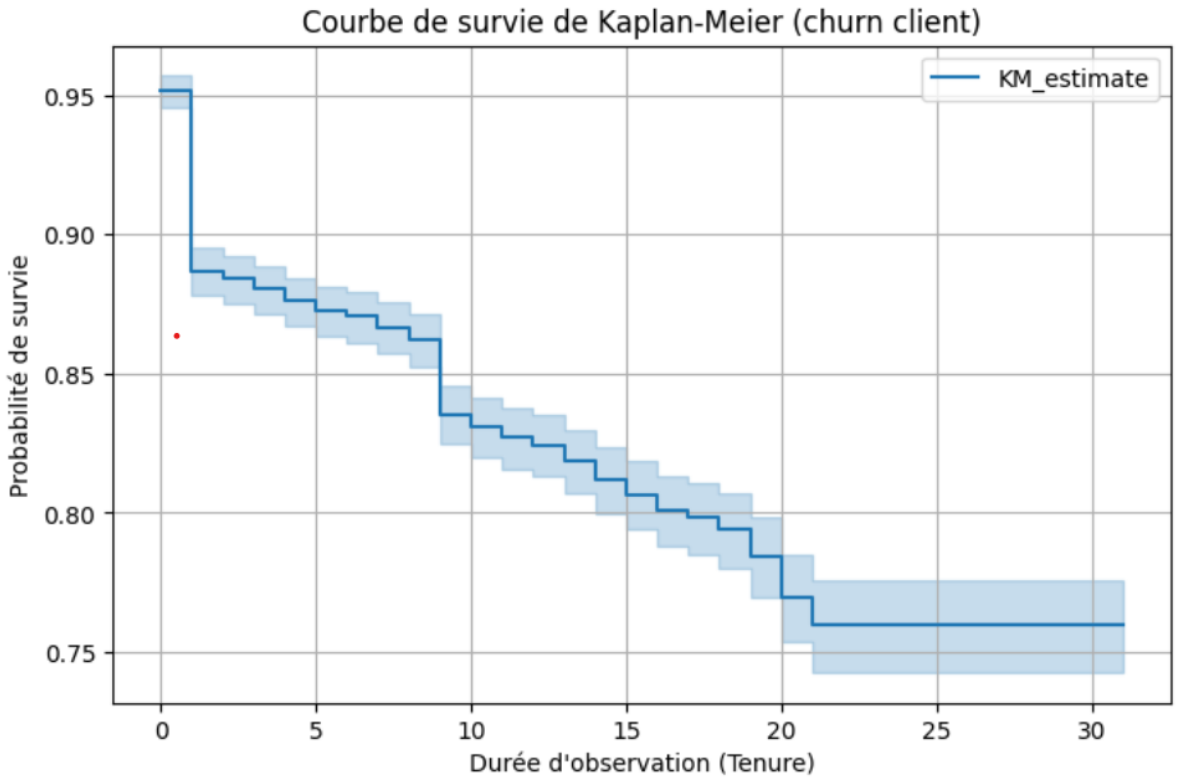
L'objectif est de compléter les prédictions obtenues par les modèles de classification par une modélisation temporelle robuste, offrant ainsi aux décideurs une vision plus opérationnelle des mécanismes de rétention et de départ des clients.

5.5. 1. Analyse exploratoire de la survie à l'aide du modèle de Kaplan-Meier

Avant d'appliquer des modèles de survie paramétriques ou semi-paramétriques, il est recommandé d'étudier la distribution empirique de la survie à l'aide de l'estimateur de Kaplan-Meier. Ce dernier est un estimateur **non paramétrique** de la fonction de survie $\hat{S}(t)$, ne reposant sur aucune hypothèse concernant la forme sous-jacente de la distribution des durées.

Résultats globaux

L'analyse de la courbe de survie globale montre qu'au tout début de la période d'observation, la probabilité de rétention chute rapidement : elle passe de **1 à moins de 0,90 dès les premiers mois**. Ensuite, en moins de **10 mois**, la probabilité descend en dessous de **0,85**, puis continue de décroître légèrement jusqu'à atteindre une valeur inférieure à **0,77 au bout de 20 mois**. On constate ainsi une décroissance progressive et régulière de la survie client, sans retour au-dessus des seuils précédemment franchis.



Effet de la variable Complain

L'estimation stratifiée selon la variable *complain* met en évidence un comportement différencié :

- Pour les clients qui **ont émis une plainte (complain = 1)**, la probabilité de survie chute brutalement à **moins de 0,75 en moins de 10 mois**.
- À l'inverse, pour les clients **n'ayant jamais porté plainte (complain = 0)**, la courbe de survie reste relativement stable et se maintient **au-dessus de 0,85 tout au long de la période observée**.

Cette observation confirme que l'insatisfaction exprimée via une plainte constitue un signal fort de risque de départ.

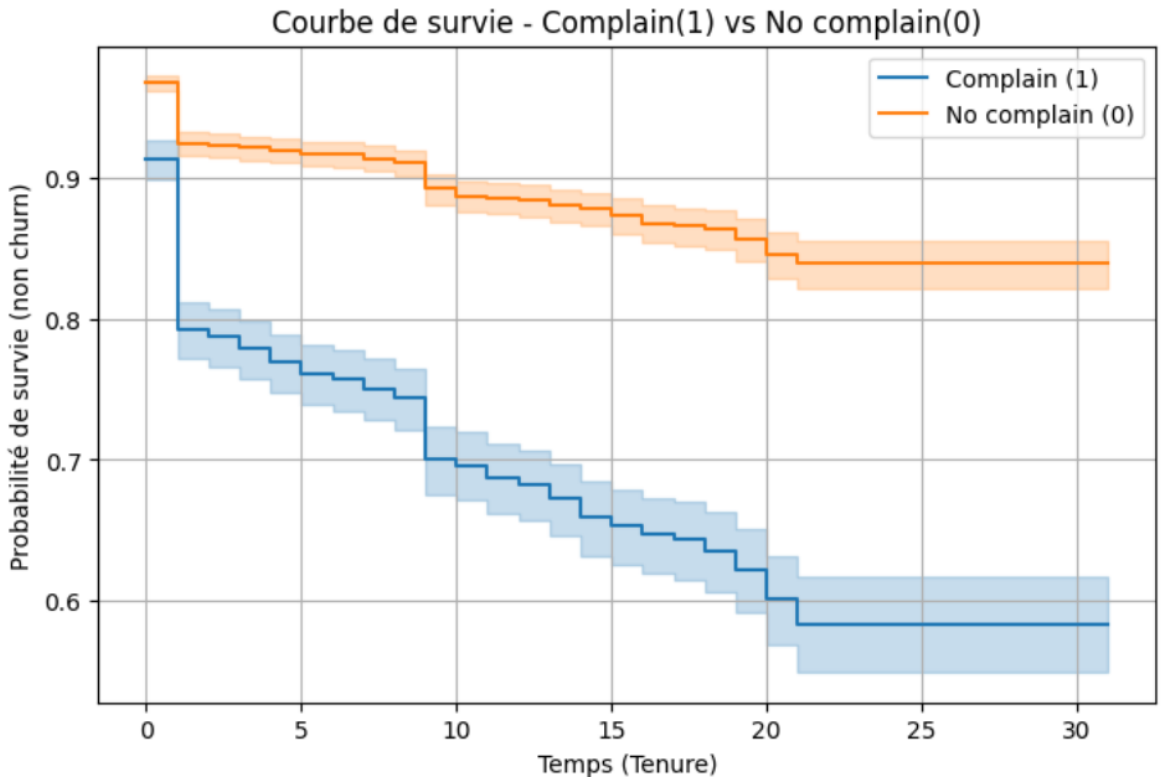


Figure X : Estimation de Kaplan-Meier – stratification par complain)

Effet de la variable *Segment*

L'estimation réalisée selon le segment client révèle également des différences notables :

- Le **segment 0** apparaît nettement plus stable, avec une survie toujours supérieure à celle des autres segments et se maintenant sur des niveaux élevés.
- En revanche, pour le **segment 1**, la probabilité de survie diminue plus rapidement, chutant en dessous de **0,80 dès le 10e mois**.

Cette analyse montre ainsi que certains profils de clients présentent une vulnérabilité accrue au churn, ce qui justifie un ciblage différencié des actions de fidélisation.

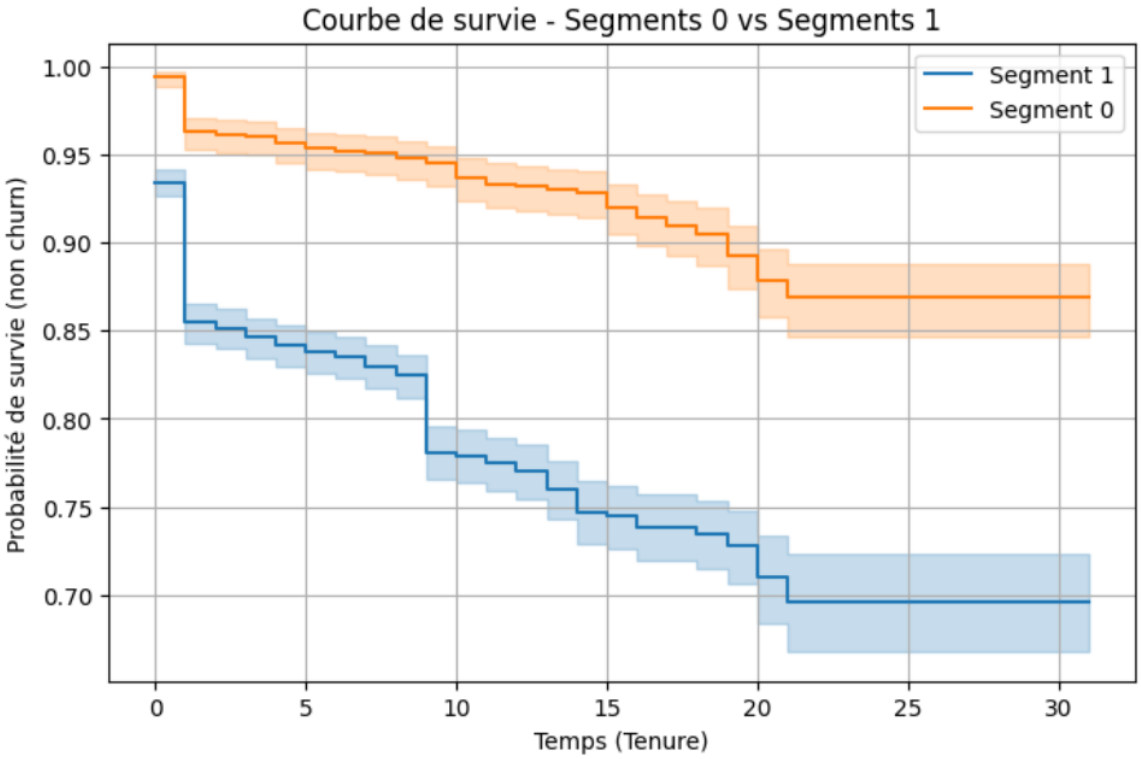


Figure : Estimation de Kaplan-Meier – stratification par segment

Cependant, bien que l'estimation de Kaplan-Meier fournisse une vision descriptive et intuitive de la probabilité de survie des clients selon différentes variables explicatives, elle reste limitée car elle ne permet pas de mesurer l'effet simultané de plusieurs covariables ni de quantifier leur impact sur le risque de départ. Pour dépasser cette limite, il est nécessaire de recourir à des modèles semi-paramétriques et paramétriques, en particulier le modèle de régression de Cox Proportionnel des Risques (CoxPH) et le modèle Accéléré du Temps de Défaillance de Weibull (Weibull AFT), qui offrent une interprétation plus fine et une meilleure capacité prédictive du phénomène de churn.

5.5.2 Application du modèle de Cox Proportionnel (CoxPH)

Afin d'intégrer la dimension temporelle dans la modélisation du churn, nous avons estimé un modèle semi-paramétrique de Cox Proportionnel (CoxPH) à partir de l'ensemble de données de survie. L'ajustement a été réalisé en incluant un ensemble de covariables pertinentes liées à la satisfaction client, aux comportements d'achat et aux préférences de consommation, dans le but de quantifier leur effet sur le risque instantané de résiliation.

Vérification de l'hypothèse des risques proportionnels

L'hypothèse centrale du modèle de Cox réside dans la **proportionnalité des risques** au cours du temps. Afin de tester cette hypothèse, nous avons appliqué le **test de Schoenfeld** sur l'ensemble des covariables.

Les résultats (Tableau X) montrent que la majorité des variables respectent l'hypothèse des risques proportionnels (p-value > 0.05), garantissant ainsi la validité globale du modèle. Toutefois, quelques covariables – notamment **HourSpendOnApp**, **PreferedOrderCat_Mobile Phone** et **DaySinceLastOrder** – présentent des p-values significatives (< 0.05), suggérant une violation locale de cette hypothèse. Néanmoins, au vu de leur poids relatif et de la bonne robustesse globale du modèle (concordance de 0.82), il a été jugé pertinent de conserver l'ensemble des covariables tout en gardant à l'esprit cette limite.

test_statistic	p	-log2(p)	
CashbackAmount	0.04	0.84	0.26
CityTier	3.35	0.07	3.90
Complain	2.17	0.14	2.83
CouponUsed	0.18	0.67	0.58
DaySinceLastOrder	5.07	0.02	5.36
Gender	1.72	0.19	2.40
HourSpendOnApp	9.08	<0.005	8.60
MaritalStatus_Single	0.29	0.59	0.75
NumberOfAddress	3.14	0.08	3.71
NumberOfDeviceRegistered	1.13	0.29	1.80
OrderAmountHikeFromlastYear	1.07	0.30	1.73
OrderCount	0.54	0.46	1.11
PreferedOrderCat_Laptop & Accessory	4.09	0.04	4.53
PreferedOrderCat_Mobile	0.70	0.40	1.31
PreferedOrderCat_Mobile Phone	8.16	<0.005	7.87
PreferredLoginDevice_Mobile Phone	0.98	0.32	1.63
PreferredLoginDevice_Phone	3.04	0.08	3.62
PreferredPaymentMode_COD	0.92	0.34	1.57
PreferredPaymentMode_Credit Card	0.22	0.64	0.64
PreferredPaymentMode_Debit Card	2.20	0.14	2.86
PreferredPaymentMode_E wallet	0.01	0.92	0.12
PreferredPaymentMode_UPI	0.02	0.89	0.16
SatisfactionScore	0.13	0.71	0.49
Segment	0.05	0.82	0.29
WarehouseToHome	0.04	0.83	0.26

Tableau X

Caractéristiques techniques du modèle CoxPH

Le modèle de Cox Proportionnel (CoxPH) a été estimé à l'aide de l'algorithme implémenté dans la librairie *lifelines*. La variable de durée retenue correspond au **Tenure** (ancienneté du client), tandis que l'événement étudié est le **Churn** (1 si le client a résilié, 0 sinon).

L'ajustement a porté sur **5 630 observations**, dont **939 événements observés**, soit un taux de churn d'environ **16,7 %**. Le nombre d'événements étant suffisant, l'estimation bénéficie d'une bonne stabilité statistique.

Une pénalisation **L2 (ridge)** de faible intensité (**penalizer = 0.1**) a été introduite dans l'optimisation. Ce choix vise à **réduire le risque de sur-ajustement** en présence d'un nombre important de covariables potentiellement corrélées, sans pour autant éliminer des variables explicatives (aucune régularisation L1 n'a été appliquée, *l1 ratio = 0.0*).

La fonction de risque de base a été estimée selon la méthode de **Breslow**, qui constitue l'approche de référence dans le cas de données groupées. La **log-vraisemblance partielle** du modèle atteint une valeur de **-7351.38**. Si cette statistique n'est pas directement interprétable, elle constitue une mesure interne de qualité d'ajustement et permet de comparer des variantes du modèle sur des bases objectives.

Ces éléments garantissent la robustesse de l'estimation et servent de fondement à l'interprétation des coefficients et Hazard Ratios présentés dans la section suivante.

Résultats CoxPH:

model	lifelines.CoxPHFitter
duration col	'Tenure'
event col	'Churn'
penalizer	0.1
l1 ratio	0.0
baseline estimation	breslow
number of observations	5630
number of events observed	939
partial log-likelihood	-7351.38

Interprétation des résultats du modèle

L'ajustement du modèle CoxPH sur **5 630 observations** (dont **939 événements de churn**) fournit des résultats statistiquement significatifs et riches d'enseignements. L'indice de

concordance (**C-index = 0.82**) témoigne d'une excellente capacité discriminante, confirmant la pertinence du modèle pour prédire le moment de l'attrition.

coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)	
CityTier	0.20	1.23	0.03	0.15	0.26	1.16	1.30	0.00	6.77	<0.005	36.20
WarehouseToHome	0.02	1.02	0.00	0.01	0.02	1.01	1.02	0.00	5.20	<0.005	22.24
Gender	0.10	1.11	0.05	0.00	0.21	1.00	1.23	0.00	1.97	0.05	4.35
HourSpendOnApp	0.01	1.01	0.04	-0.07	0.08	0.93	1.09	0.00	0.19	0.85	0.23
NumberOfDeviceRegistered	0.18	1.20	0.03	0.13	0.23	1.14	1.26	0.00	6.71	<0.005	35.56
SatisfactionScore	0.12	1.13	0.02	0.08	0.16	1.09	1.17	0.00	6.30	<0.005	31.63
NumberOfAddress	0.03	1.03	0.01	0.01	0.05	1.01	1.05	0.00	2.55	0.01	6.54
Complain	0.79	2.19	0.05	0.68	0.89	1.97	2.44	0.00	14.59	<0.005	157.70
OrderAmountHikeFromlastYear	-0.01	0.99	0.01	-0.02	0.01	0.98	1.01	0.00	-0.79	0.43	1.22
CouponUsed	0.02	1.02	0.02	-0.01	0.05	0.99	1.06	0.00	1.49	0.14	2.88
OrderCount	0.01	1.01	0.01	-0.01	0.03	0.99	1.03	0.00	0.68	0.50	1.00
DaySinceLastOrder	-0.00	1.00	0.00	-0.01	0.01	0.99	1.01	0.00	-0.07	0.94	0.08
CashbackAmount	-0.00	1.00	0.00	-0.01	-0.00	0.99	1.00	0.00	-6.46	<0.005	33.11
PreferredLoginDevice_Mobile Phone	-0.23	0.79	0.06	-0.34	-0.12	0.71	0.89	0.00	-4.03	<0.005	14.12
PreferredLoginDevice_Phone	0.01	1.01	0.07	-0.12	0.14	0.89	1.15	0.00	0.12	0.90	0.15
PreferredPaymentMode_COD	0.34	1.41	0.10	0.15	0.54	1.16	1.71	0.00	3.50	<0.005	11.09
PreferredPaymentMode_Credit Card	-0.14	0.87	0.07	-0.27	-0.01	0.76	0.99	0.00	-2.06	0.04	4.65
PreferredPaymentMode_Debit Card	-0.07	0.93	0.06	-0.19	0.05	0.83	1.05	0.00	-1.16	0.25	2.02
PreferredPaymentMode_E wallet	0.22	1.25	0.09	0.05	0.40	1.05	1.49	0.00	2.54	0.01	6.48
PreferredPaymentMode_UPI	0.02	1.02	0.10	-0.18	0.23	0.84	1.25	0.00	0.23	0.82	0.29
PreferedOrderCat_Laptop & Accessory	-0.34	0.71	0.06	-0.46	-0.22	0.63	0.81	0.00	-5.43	<0.005	24.06
PreferedOrderCat_Mobile	0.43	1.54	0.08	0.27	0.59	1.31	1.81	0.00	5.19	<0.005	22.21
PreferedOrderCat_Mobile Phone	0.39	1.48	0.07	0.26	0.53	1.30	1.70	0.00	5.74	<0.005	26.65
MaritalStatus_Single	0.56	1.75	0.05	0.45	0.66	1.57	1.94	0.00	10.38	<0.005	81.39
Segment	0.41	1.51	0.07	0.27	0.55	1.31	1.74	0.00	5.71	<0.005	26.37

TABLEAU X

Variables sociodémographiques et structurelles

- **CityTier (HR = 1.23, p < 0.005)** : les clients provenant de villes de rang plus élevé présentent un risque accru de churn.
- **Segment (HR = 1.51, p < 0.005)** : appartenir au segment 1 multiplie le risque de résiliation par 1.51.
- **MaritalStatus_Single (HR = 1.75, p < 0.005)** : les célibataires sont significativement plus enclins à résilier.

Satisfaction et engagement

- **SatisfactionScore (HR = 1.13, p < 0.005)** : de façon paradoxale, une hausse du score de satisfaction est associée à un risque plus élevé de churn, ce qui pourrait refléter un biais de perception ou une clientèle plus exigeante.
- **Complain (HR = 2.19, p < 0.005)** : les clients ayant déposé une plainte sont plus de deux fois plus susceptibles de résilier, soulignant l'impact majeur de l'insatisfaction.

Comportements transactionnels et préférences

- **Modes de paiement :**
  - Paiement à la livraison (COD) : +41% de risque (HR = 1.41, p < 0.005).
  - Carte de crédit : effet protecteur (HR = 0.87, p = 0.04).
  - E-wallet : +25% de risque (HR = 1.25, p = 0.01).
- **Catégories de produits :**
  - Laptop & Accessory → protecteur (HR = 0.71, p < 0.005).
  - Mobile (HR = 1.54) et Mobile Phone (HR = 1.48) → fortement associés à un risque accru.
- **Autres variables :**
  - Nombre d'adresses enregistrées (HR = 1.03, p = 0.01) et nombre de dispositifs connectés (HR = 1.20, p < 0.005) augmentent le risque de churn, traduisant possiblement une relation client plus complexe.
  - CashbackAmount (HR ≈ 1.00 mais coefficient négatif, p < 0.005) a un effet protecteur très faible mais significatif, suggérant que les politiques de cashback peuvent contribuer à la fidélisation.

La figure X présente la visualisation des coefficients estimés pour l'ensemble des covariables du modèle CoxPH, accompagnés de leurs intervalles de confiance à 95 %. Sur ce graphique, la **ligne verticale à 0** représente le seuil neutre : un coefficient supérieur à 0 indique que la variable **augmente le risque de churn**, tandis qu'un coefficient inférieur à 0 signale un **effet protecteur**.

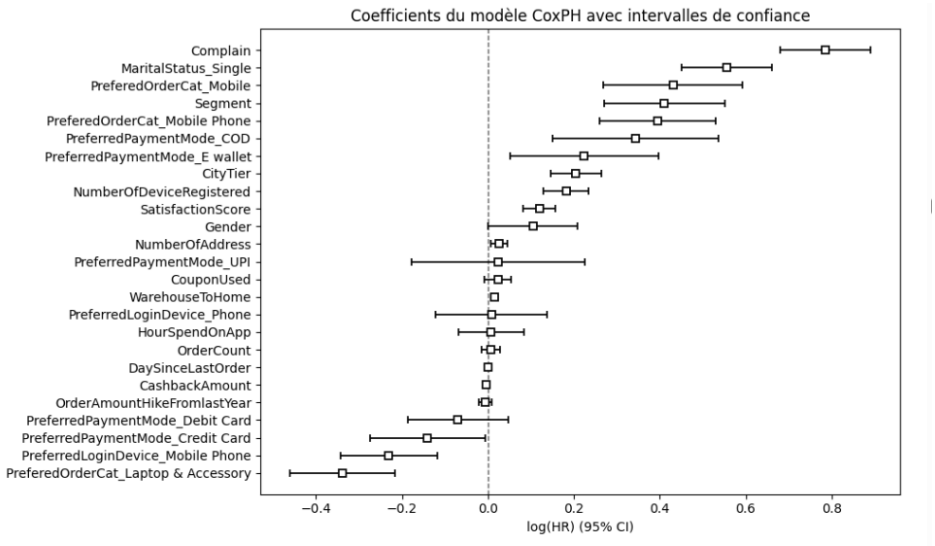
Les intervalles de confiance permettent d'évaluer la significativité statistique. Les variables dont l'intervalle ne chevauche pas 0 exercent un effet significatif sur le risque de résiliation. On observe ainsi que des facteurs tels que **Complain**, **MaritalStatus_Single**, **Segment**, **CityTier**, **NumberOfDeviceRegistered** et certaines catégories de produits (*Mobile* et *Mobile Phone*) ont

un coefficient positif significatif, traduisant une **augmentation substantielle du risque de churn**.

À l'inverse, certaines covariables présentent des coefficients négatifs et significatifs. C'est le cas de **CashbackAmount**, **PreferredLoginDevice_Mobile Phone**, **PreferredPaymentMode_Credit Card** et **PreferedOrderCat_Laptop & Accessory**, qui apparaissent comme des facteurs **protecteurs**, réduisant la probabilité de résiliation.

Enfin, d'autres variables telles que **OrderCount**, **DaySinceLastOrder** ou **OrderAmountHikeFromlastYear** ont des intervalles de confiance incluant 0, ce qui suggère que leur effet sur le churn **n'est pas statistiquement significatif** dans ce modèle.

Cette visualisation offre ainsi une synthèse intuitive des covariables influençant le churn et permet de **hiérarchiser les leviers stratégiques** pour la fidélisation client, en fonction de leur impact relatif et de leur significativité.



Limites et Transition

Bien que le modèle de Cox Proportionnel des Risques constitue une approche de référence en analyse de survie en raison de sa flexibilité et de l'absence d'hypothèse explicite sur la forme de la fonction de risque de base, il présente néanmoins certaines limites. La principale contrainte réside dans l'hypothèse de proportionnalité des risques, supposant que le rapport des risques entre individus reste constant dans le temps. Or, cette hypothèse peut être violée dans des contextes où les effets des covariables évoluent au fil du temps, entraînant ainsi un biais dans les estimations. De plus, le modèle de Cox ne permet pas une estimation explicite de la fonction de survie sans recourir à une estimation supplémentaire de la fonction de risque de base, ce qui peut limiter l'interprétation directe dans certains cas appliqués.

Afin de surmonter ces limites, en particulier lorsque l'on souhaite disposer d'une forme paramétrique explicite de la fonction de survie et du risque, il est pertinent de recourir à un modèle de type **Accelerated Failure Time (AFT)**. Parmi les distributions possibles, le modèle de Weibull AFT s'avère particulièrement adapté car il permet à la fois de modéliser directement le temps de survie et d'obtenir une fonction de survie analytique. Ce cadre paramétrique fournit non seulement une meilleure interprétabilité des coefficients en termes de facteur d'accélération

du temps de survie, mais également une estimation robuste même en présence de jeux de données de taille modeste.

5.5.3 Application du modèle Weibull AFT

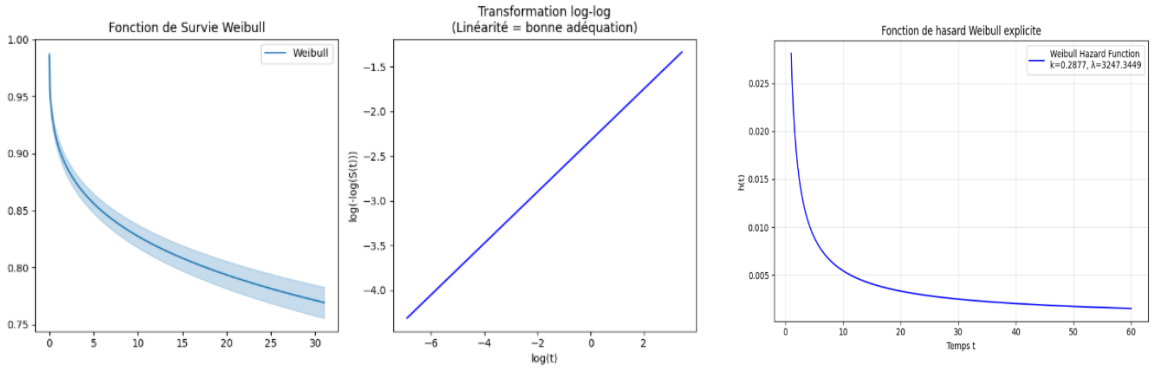
Afin de dépasser les limites du modèle de Cox, notamment son hypothèse de proportionnalité des risques, nous avons appliqué un modèle paramétrique de type **Accelerated Failure Time (AFT)** basé sur la loi de **Weibull**. Ce choix se justifie par la flexibilité de ce modèle, qui permet de modéliser directement le temps de survie et d'interpréter les covariables comme des facteurs d'accélération ou de décélération de la durée de vie des clients.

a) Vérification de l'adéquation au modèle Weibull

Avant d'estimer le modèle, une vérification de l'adéquation de la distribution de Weibull a été réalisée. Les paramètres estimés par ajustement simple sont :

- **Paramètre de forme (k = ρ)** : `0.2877`
- **Paramètre d'échelle (λ)** : `3247.34`

Les diagnostics graphiques confirment la pertinence de cette hypothèse. En particulier, la transformation **log(-log(S(t)))** en fonction de **log(t)** montre une relation approximativement linéaire, ce qui valide l'adéquation des données à une loi de Weibull. De plus, la fonction de risque estimée présente une dynamique cohérente avec le comportement attendu des durées de vie en contexte de churn.



b) Ajustement du modèle Weibull AFT

Le modèle Weibull AFT a ensuite été ajusté sur l'échantillon, composé de **5630 observations**, dont **939 événements de churn observés**. Les principaux indicateurs de qualité d'ajustement sont encourageants :

Concordance	0.82
AIC	5696.84
log-likelihood ratio test	1263.47 on 25 df
-log2(p) of ll-ratio test	831.42

- **Indice de concordance : 0,8151**, révélant une excellente capacité discriminante du modèle ;
- **AIC : 5696,84**, indiquant une parcimonie raisonnable de l'ajustement ;
- **Log-likelihood ratio test : 1263,47 (25 ddl)**, hautement significatif (**-log2(p) = 831,42**), ce qui confirme la contribution des covariables incluses.

c) Interprétation des résultats

Les résultats détaillés des coefficients (Tableau ci-dessous) offrent plusieurs enseignements majeurs :

1. **Variables significatives et impact fort :**
   - La variable **Complain** exerce l'effet le plus marqué : les clients ayant exprimé une plainte présentent un temps de survie considérablement réduit, avec un facteur d'accélération exp(coef) ≈ 0,03 ($p < 0,005$).
   - Le **segment** influence également fortement le churn : les individus du segment 1 ont une probabilité accrue de départ rapide (exp(coef) ≈ 0,07).
   - Le **statut marital (Single)** joue un rôle protecteur : exp(coef) ≈ 0,09, traduisant une survie plus longue que les autres profils.
2. **Effets modérés mais significatifs :**
   - Le **CityTier** (exp(coef) ≈ 0,33) indique que résider dans un certain type de ville réduit notablement la durée de survie.
   - Le **nombre d'adresses** et le **nombre d'appareils enregistrés** présentent des effets négatifs significatifs (exp(coef) ≈ 0,81 et 0,38 respectivement), suggérant une instabilité comportementale accrue.
   - Le **SatisfactionScore** réduit fortement le risque de churn : un coefficient négatif (exp(coef) ≈ 0,54) confirme que la satisfaction est un facteur central de rétention.
3. **Effets positifs (accélérateurs de churn) :**
   - Le choix de la catégorie **Laptop & Accessory** comme préférence d'achat multiplie par plus de 7 le risque relatif de churn (exp(coef) ≈ 7,27).
   - Le **PreferredLoginDevice_Mobile Phone** et **Phone** augmentent également la vitesse de churn (exp(coef) ≈ 2,40 et 1,82).
4. **Variables non significatives :**

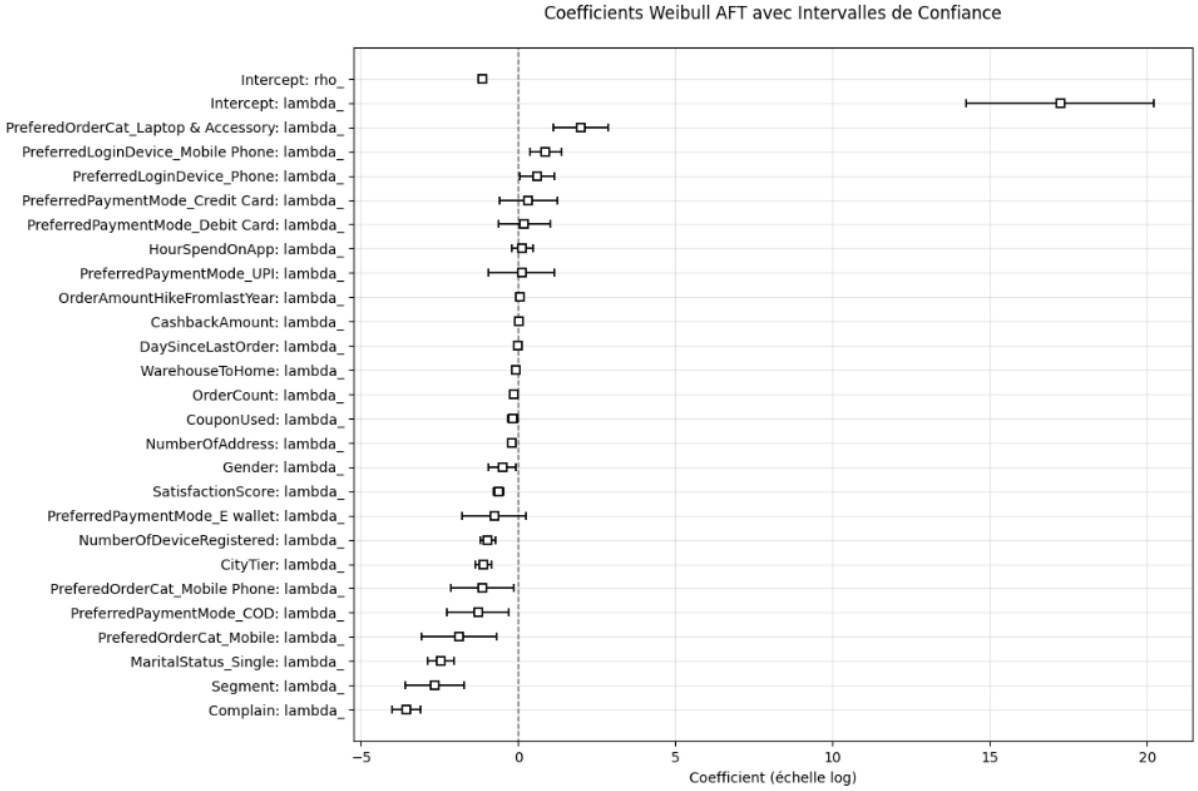
- Certaines modalités, comme l'utilisation de la **carte de crédit** ou du **paiement via UPI**, ne présentent pas d'effet statistiquement significatif, suggérant que leur influence est marginale ou confondue avec d'autres facteurs.

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)	
lambda_	CashbackAmount	0.02	1.02	0.00	0.01	0.03	1.01	1.03	0.00	4.77	<0.005	19.02
	CityTier	-1.11	0.33	0.13	-1.38	-0.85	0.25	0.43	0.00	-8.30	<0.005	53.09
	Complain	-3.56	0.03	0.23	-4.01	-3.11	0.02	0.04	0.00	-15.58	<0.005	179.40
	CouponUsed	-0.16	0.85	0.08	-0.32	-0.01	0.73	0.99	0.00	-2.06	0.04	4.68
	DaySinceLastOrder	-0.02	0.98	0.01	-0.05	0.00	0.95	1.00	0.00	-1.71	0.09	3.52
	Gender	-0.51	0.60	0.22	-0.93	-0.08	0.39	0.92	0.00	-2.35	0.02	5.74
	HourSpendOnApp	0.12	1.13	0.17	-0.22	0.46	0.80	1.58	0.00	0.70	0.48	1.05
	MaritalStatus_Single	-2.46	0.09	0.22	-2.88	-2.03	0.06	0.13	0.00	-11.34	<0.005	96.52
	NumberOfAddress	-0.21	0.81	0.04	-0.28	-0.14	0.75	0.87	0.00	-5.52	<0.005	24.77
	NumberOfDeviceRegistered	-0.97	0.38	0.12	-1.21	-0.73	0.30	0.48	0.00	-7.97	<0.005	49.17
	OrderAmountHikeFromlastYear	0.05	1.05	0.03	-0.01	0.11	0.99	1.11	0.00	1.78	0.07	3.75
	OrderCount	-0.13	0.88	0.06	-0.24	-0.02	0.78	0.98	0.00	-2.29	0.02	5.52
	PreferedOrderCat_Laptop & Accessory	1.98	7.27	0.44	1.12	2.85	3.06	17.28	0.00	4.49	<0.005	17.08
	PreferedOrderCat_Mobile	-1.87	0.15	0.61	-3.06	-0.69	0.05	0.50	0.00	-3.09	<0.005	8.97
	PreferedOrderCat_Mobile Phone	-1.13	0.32	0.51	-2.13	-0.12	0.12	0.88	0.00	-2.20	0.03	5.18
	PreferredLoginDevice_Mobile Phone	0.88	2.40	0.26	0.37	1.38	1.45	3.99	0.00	3.40	<0.005	10.54
	PreferredLoginDevice_Phone	0.60	1.82	0.28	0.04	1.15	1.05	3.17	0.00	2.12	0.03	4.87
	PreferredPaymentMode_COD	-1.28	0.28	0.50	-2.25	-0.30	0.11	0.74	0.00	-2.56	0.01	6.59
	PreferredPaymentMode_Credit Card	0.32	1.38	0.47	-0.60	1.25	0.55	3.50	0.00	0.68	0.49	1.02
	PreferredPaymentMode_Debit Card	0.19	1.21	0.42	-0.62	1.01	0.54	2.75	0.00	0.46	0.64	0.64
	PreferredPaymentMode_E wallet	-0.76	0.47	0.52	-1.77	0.25	0.17	1.29	0.00	-1.47	0.14	2.83
	PreferredPaymentMode_UPI	0.11	1.12	0.54	-0.94	1.16	0.39	3.20	0.00	0.20	0.84	0.25
	SatisfactionScore	-0.62	0.54	0.08	-0.78	-0.47	0.46	0.63	0.00	-7.71	<0.005	46.22
	Segment	-2.65	0.07	0.47	-3.58	-1.73	0.03	0.18	0.00	-5.61	<0.005	25.59
	WarehouseToHome	-0.08	0.92	0.01	-0.11	-0.06	0.90	0.94	0.00	-6.93	<0.005	37.82
	Intercept	17.23	3.05e+07	1.53	14.24	20.23	1.53e+06	6.08e+08	0.00	11.29	<0.005	95.74
rho_	Intercept	-1.13	0.32	0.03	-1.19	-1.07	0.30	0.34	0.00	-37.47	<0.005	1018.23

La figure présente les **coefficients estimés du modèle Weibull AFT** avec leurs **intervalles de confiance à 95 %**. Chaque barre horizontale représente l'effet d'une covariable sur le **log du temps jusqu'au churn** :

- Une **barre entièrement à droite de zéro** (coef > 0) indique que l'augmentation de la variable **prolonge le temps de survie**, donc réduit le risque de churn.
- Une **barre entièrement à gauche de zéro** (coef < 0) montre que la variable **accélère le churn**, diminuant la durée de fidélité.
- Les barres traversant zéro signalent des effets **non significatifs**, où l'impact sur le temps jusqu'au churn est incertain.

Cette visualisation permet de **hiérarchiser rapidement l'importance des facteurs** influençant la fidélité des clients et de détecter ceux qui ont un effet protecteur ou aggravant sur le churn.



5.5.4 Comparaison des performances des modèles CoxPH et Weibull AFT

L'évaluation de la performance prédictive des modèles de survie repose généralement sur deux familles de métriques complémentaires :

- **L'indice de concordance (C-index)**, proposé par Harrell, mesure la capacité d'un modèle à bien ordonner les temps de survie prédits par rapport aux observations réelles. Sa valeur varie entre 0.5 (équivalent au hasard) et 1 (parfaite discrimination).
- **Le Brier Score** (Graf et al., 1999), qui quantifie l'exactitude des probabilités de survie estimées dans le temps. Pour un instant donné ttt, il s'agit d'une mesure de l'erreur quadratique entre la survie prédite et l'événement observé. Le score est ensuite corrigé par pondération inverse de la probabilité de censure (IPCW), ce qui permet une

estimation non biaisée en présence de données censurées. Plus le Brier Score est faible, meilleure est la calibration du modèle.

Dans notre étude, nous avons comparé les modèles **Cox Proportional Hazards (CoxPH)** et **Accelerated Failure Time (AFT) basé sur une loi de Weibull**. Les résultats sont synthétisés dans les tableaux ci-dessous.

Résultats des indices de concordance

Modèle	C-index
CoxPH	0.8151
Weibull AFT	0.8211

Les deux modèles présentent un pouvoir discriminant élevé, avec des indices de concordance voisins de 0.82. Le modèle Weibull AFT obtient un score légèrement supérieur, indiquant une capacité marginalement meilleure à classer correctement les durées de survie.

Résultats des Brier Scores corrigés (IPCW)

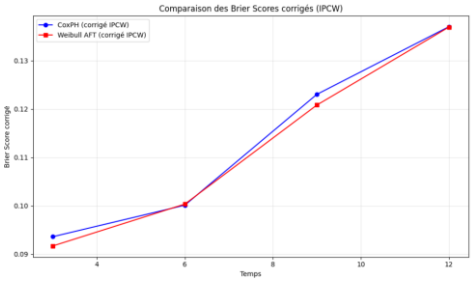
Temps (mois)	CoxPH	Weibull AFT
3	0.0936	0.0917
6	0.1001	0.1003
9	0.1231	0.1208
12	0.1370	0.1370

L'évolution temporelle des Brier Scores montre que :

- à **court terme (3 mois)**, le Weibull AFT présente un léger avantage avec un score plus bas (meilleure calibration),
- à **moyen terme (6 mois)**, les performances des deux modèles sont quasiment identiques,
- à **9 mois**, le Weibull AFT reprend un avantage en calibration,
- à **long terme (12 mois)**, les performances convergent de nouveau.

Ces résultats soulignent que, sur notre jeu de données, le modèle Weibull AFT est au moins aussi performant que le CoxPH en termes de calibration probabiliste, tout en offrant une **modélisation paramétrique explicite** de la fonction de survie, avantageuse pour la prédiction de probabilités individuelles à différents horizons.

Une représentation graphique des Brier Scores au cours du temps est proposée à la **Figure X**, afin de visualiser directement la dynamique comparative des performances des deux modèles.



Integrated Brier Score (IBS) jusqu'à 12 mois

L'**Integrated Brier Score (IBS)** correspond à l'aire sous la courbe des Brier Scores, normalisée par l'horizon de temps considéré τ:

$$IBS(0,\tau) = \frac{1}{\tau}\int_0^{\tau} BS(t)dt$$

Il fournit donc une **mesure globale et synthétique** de la calibration d'un modèle de survie sur toute la période [0,τ], plutôt qu'à un instant isolé.

Le résultat est donné par ce tableau **:**

Modèle	**IBS (0–12 mois)**
CoxPH	0.0846
Weibull AFT	0.0839

Dans notre cas :

- le **Weibull AFT** obtient un IBS légèrement inférieur à celui du CoxPH, ce qui confirme son **meilleur calibrage global**.
- cette métrique résume efficacement l'information fournie par les Brier Scores individuels à 3, 6, 9 et 12 mois, offrant ainsi un indicateur synthétique très utile pour comparer des modèles de survie dans un contexte appliqué, comme la prédiction du churn client.

L'IBS est particulièrement pertinent dans le suivi des clients sur plusieurs mois, car il reflète la capacité d'un modèle à fournir des probabilités précises et utilisables à tout moment. Une calibration précise permet à l'entreprise d'**anticiper le churn** et de mettre en place des actions de fidélisation adaptées, en priorisant les clients les plus à risque selon les probabilités prédictives du modèle.

5.5.5 Discussion et conclusion

L'analyse de survie appliquée au churn a permis de comparer plusieurs approches complémentaires. La méthode **Kaplan-Meier**, non paramétrique, offre une première vision descriptive et intuitive des courbes de survie entre différents groupes de clients. Elle est précieuse pour mettre en évidence des différences globales de comportement (par exemple selon le segment ou le niveau de satisfaction). Cependant, son principal inconvénient réside dans son incapacité à intégrer plusieurs covariables simultanément : elle ne permet pas de quantifier précisément l'effet de facteurs multiples sur le risque de churn.

Le modèle **Cox Proportionnal Hazards (CoxPH)** a ensuite apporté une avancée significative, en permettant d'évaluer l'impact de chaque covariable sur le risque instantané de départ, via les coefficients interprétables en termes de fonction de hasard. De plus, sa souplesse repose sur le fait qu'il ne spécifie pas la forme de la fonction de base du risque. Toutefois, cette flexibilité a une contrepartie : CoxPH ne fournit pas directement le temps de survie attendu, mais plutôt un

rapport de hasards relatif. De plus, l'hypothèse de proportionnalité des risques doit être vérifiée, sous peine de conclusions biaisées.

Le modèle **Weibull AFT (Accelerated Failure Time)** se distingue par sa nature paramétrique. Contrairement au CoxPH, il permet une interprétation directe en termes de temps de survie : les coefficients indiquent si une variable accélère ou ralentit le moment du churn. Dans notre cas, les diagnostics statistiques ont confirmé l'adéquation de la loi de Weibull, validant l'usage de ce modèle. Ses atouts sont doubles : d'une part, il enrichit la compréhension en donnant un horizon temporel plus concret (par exemple, combien de temps un client reste avant de quitter en fonction de ses caractéristiques) ; d'autre part, il permet de calculer **directement une probabilité de survie à n'importe quel instant**, ce qui constitue un outil opérationnel pour la fidélisation. En pratique, cela signifie que l'on peut estimer la probabilité qu'un client soit encore actif dans 3, 6 ou 12 mois, et ajuster les actions marketing ou relationnelles en conséquence.

Néanmoins, cette approche exige une hypothèse forte (ici la loi de Weibull) et reste sensible à une mauvaise spécification du modèle. Une telle contrainte impose une vigilance méthodologique accrue. Mais dans notre étude, la validité confirmée du modèle garantit la solidité des prédictions.

En conclusion, l'approche Kaplan-Meier reste utile pour une visualisation descriptive, CoxPH permet de mesurer des effets relatifs de manière flexible, et Weibull AFT offre une interprétation prédictive directement exploitable pour des systèmes de fidélisation. Dans le cadre d'un dispositif décisionnel, l'AFT apparaît donc comme l'outil le plus adapté, puisqu'il combine interprétabilité, précision temporelle et potentiel opérationnel.

5.6 Estimation de Customer Lifetime Value

Introduction

Après avoir analysé les comportements clients à travers des approches de **classification** et de **survie**, il est pertinent d'étendre la réflexion vers une dimension plus stratégique : la **valeur vie client (Customer Lifetime Value, CLV)**.
En effet, si la classification permet d'identifier les clients les plus susceptibles de **churner** et que l'analyse de survie fournit une estimation de la **durée de rétention**, aucune de ces deux approches ne répond directement à une question essentielle pour les entreprises : *« Quelle est la valeur économique attendue d'un client sur une période donnée ? »*

C'est précisément l'avantage de l'estimation du CLV.
En intégrant non seulement la probabilité de survie du client (sa rétention), mais aussi une approximation de son **potentiel de dépense futur**, le CLV offre une vision plus complète et plus opérationnelle.
Ainsi, il ne s'agit plus uniquement de prédire si un client va rester ou partir, mais d'anticiper la **contribution financière associée à sa présence** dans le portefeuille de l'entreprise.

Cette estimation, même sur un jeu de données agrégé et limité (par exemple via la variable *CashbackAmount* comme proxy des montants dépensés), enrichit l'étude précédente. Elle permet de passer d'une analyse centrée sur le comportement client (durée, churn, satisfaction)

à une analyse orientée vers la **décision économique et marketing**, où l'enjeu n'est plus seulement de retenir les clients, mais de **prioriser ceux qui génèrent le plus de valeur**.

5.6.1 Importance stratégique du Customer Lifetime Value (CLV)

La **valeur vie client** (Customer Lifetime Value, CLV) peut être définie comme la valeur actualisée des profits nets générés par un client au cours de sa relation avec l'entreprise. Autrement dit, elle traduit la contribution financière d'un client, en intégrant à la fois ses comportements d'achat passés et son potentiel futur.
Comme le rappellent de nombreux chercheurs, *« la notion de valeur vie d'un client est largement acceptée, tant par les chercheurs que par les praticiens du monde des affaires »* (Jain, D.C.; Singh, S.S. Customer lifetime value research in marketing: A review and future directions. J. Interact. Mark. 2002).

De la logique produit à la logique client

Historiquement, les entreprises considéraient leurs transactions avec les clients comme des **actes isolés**. La rentabilité était recherchée principalement par la réduction des coûts, la différenciation produit et la compétitivité prix. Dans cette logique centrée sur le produit, *« une série de transactions avec un client sur une période donnée n'était pas accordée l'importance qu'elle mérite dans la formulation des stratégies »*.
Cette approche présentait une limite fondamentale : elle négligeait le rôle du client comme générateur de valeur.

Avec l'évolution du marketing relationnel et l'essor du commerce électronique, les entreprises adoptent désormais une **approche centrée sur le client**. Comme le soulignent **A. Persson et L. Ryals** (Persson, A.; Ryals, L. Customer assets and customer equity: Management and measurement issues. Mark. Theory 2010, 10, 417–436.) cette approche considère les clients comme de véritables **actifs immatériels** et met l'accent sur l'acquisition, la fidélisation et la valorisation des relations clients, constituant ainsi une base d'**avantage concurrentiel durable**.
Contrairement aux produits, qui peuvent être rapidement copiés, une clientèle fidèle et rentable constitue une ressource difficilement imitable.

CLV et fidélité : un éclairage économique

L'un des apports fondamentaux du CLV est de replacer la **fidélité client** dans une perspective économique. En effet, *« il est important de comprendre que la fidélité n'a de valeur que dans le cadre de clients rentables. La fidélité de clients non rentables n'apporte aucun avantage à une entreprise »*.

De plus, dans un contexte où *« le coût d'acquisition d'un client est plus élevé sur Internet, la rentabilité ne peut être atteinte que si ce dernier effectue de nombreux achats répétés au cours des années suivantes »* (*Reichheld et Shefter, 2000)*
Ainsi, le CLV offre une métrique fiable pour distinguer les clients stratégiques, justifiant les efforts marketing, des clients peu rentables qui peuvent être progressivement dépriorisés (Haenlein et al., 2006 [1]).

Avantages stratégiques et tactiques

Le CLV se positionne à la fois comme un outil stratégique et tactique :

- **Stratégique** : il permet d'identifier les clients les plus précieux, de segmenter la clientèle et de cibler les efforts de fidélisation à long terme.
- **Tactique** : il guide l'allocation optimale des ressources marketing à court terme, en priorisant les clients générateurs de profits (Mulhern, 1999).

Il constitue aussi un véritable **pont entre marketing et finance**, reliant directement les actions marketing à leurs impacts financiers (*Williams, C.; Williams, R. Optimizing acquisition and retention spending to maximize market share. J. Mark.Anal.* ***2015****, 3, 159–170.*). De ce fait, le CLV dépasse les analyses rétrospectives (comme le chiffre d'affaires historique par client, CP) et fournit une **perspective prospective** indispensable pour anticiper et orienter les stratégies de croissance ([*Estrella-Ramón, A.M.; Sánchez-Pérez, M.; Swinnen, G.; VanHoof, K. A marketing view of the customer value: Customer lifetime value and customer equity. S. Afr. J. Bus. Manag. 2013, 44, 47–64.*]).

CLV dans l'ère digitale et du CRM

L'**avènement d'Internet** a amplifié l'importance du CLV.
De nombreuses entreprises digitales n'ayant pas d'actifs physiques de grande valeur, *« leur évaluation correcte ne peut se faire qu'en prenant en compte la valeur de leurs actifs immatériels. Or, la valeur de leur base clients constitue l'actif immatériel le plus important »* (*Jain & Singh, 2002*).

Le CLV devient alors la métrique clé pour mesurer la valeur réelle de l'entreprise et son potentiel de croissance future.

Par ailleurs, l'ère du **marketing de masse** est progressivement remplacée par celle du **marketing ciblé et interactif**. La connaissance du CLV permet de concevoir des programmes marketing individualisés, renforçant leur efficacité et leur efficience (Jain & Singh, 2002). Dans cette optique, le CLV s'intègre pleinement aux démarches de **Customer Relationship Management (CRM)** : il permet de segmenter la clientèle selon la valeur créée, d'identifier les segments stratégiques et de mettre en place des stratégies de fidélisation adaptées ([*Kim, S.-Y.; Jung, T.-S.; Suh, E.-H.; Hwang, H.-S. Customer segmentation and strategy development based on customer lifetime value: A case study. Expert Syst. Appl. 2006, 31, 101–107*]).

Vers une gestion prédictive et proactive

Enfin, grâce aux progrès des **technologies de l'information et de la communication (TIC)**, il est désormais possible de modéliser la CLV de façon dynamique, en intégrant les habitudes d'achat, les préférences et l'évolution du comportement client.
« En analysant et en exploitant les données historiques à travers la modélisation, il devient possible de prédire les actions futures des clients et de mettre en place une gestion ciblée » (Zhuoer Ma, 2025).

Ainsi, *« en intégrant le cycle de vie client et son potentiel futur, les entreprises peuvent ajuster continuellement leurs stratégies en suivant l'évolution des comportements, garantissant la maximisation de la valeur client »* (Zhuoer Ma, 2025).

5.6.2 Préparation des données pour l'estimation du Customer Lifetime Value

Préparation des données pour la prédiction du CLV

La prédiction du Customer Lifetime Value (CLV) à l'aide des modèles BG/NBD et Gamma-Gamma nécessite des données transactionnelles détaillées (identifiant client, date des transactions et montants associés). Or, dans le cas présent, le jeu de données initial est fourni sous une forme **agrégée**, contenant des variables clients (profil, préférences, satisfaction, etc.), ainsi que des indicateurs synthétiques tels que le nombre de commandes, le nombre de jours depuis la dernière commande et le montant de cashback. Une étape de reconstruction et de transformation des données était donc indispensable afin de les rendre compatibles avec les exigences du modèle.

1. Filtrage et sélection des clients pertinents

Afin de garantir la cohérence et la représentativité des résultats, seuls les clients présentant une activité significative ont été retenus. Deux conditions de filtrage ont été appliquées :

- une ancienneté minimale de **60 jours** (soit deux mois),
- un nombre de commandes supérieur à **une unité**.

Ce choix méthodologique permet d'éliminer les clients nouvellement acquis ou très peu actifs, dont les données ne contiennent pas suffisamment d'informations temporelles pour calibrer un modèle fiable.

2. Estimation du volume transactionnel

À partir des variables disponibles, une estimation du nombre total de transactions a été réalisée. En pratique, le nombre de transactions d'un client est approché par la formule :

$$N^{\wedge}\text{transactions} \approx \text{Tenure}_{\text{mois}} \times \text{OrderCount}$$

où *Tenure* désigne la durée de la relation client exprimée en mois et *OrderCount* correspond au nombre moyen de commandes par mois observé. Cette approximation, bien que simplificatrice, permet de reconstruire une base transactionnelle cohérente avec la dynamique du client.

3. Génération des dates de transactions

Une fois le volume transactionnel estimé, les dates des transactions ont été simulées dans une fenêtre temporelle comprise entre la date de début de relation et la date de fin d'observation. Les transactions ont été espacées régulièrement, puis perturbées par un bruit aléatoire (±10 jours), de manière à reproduire une variabilité réaliste dans les comportements d'achat. Ce procédé génère une granularité temporelle indispensable pour le calcul de la récence et de l'ancienneté observée (*recency* et *T*), variables clés du modèle BG/NBD.

4. Reconstruction des montants transactionnels

La variable monétaire, indispensable pour l'estimation du CLV à l'aide du modèle Gamma-Gamma, a été reconstruite à partir des montants de cashback reportés dans les données initiales. L'hypothèse adoptée repose sur le fait que le cashback représente un pourcentage du panier d'achat, variable selon la catégorie de produit. Par exemple, les articles de mode bénéficient d'un taux de cashback plus élevé que l'électronique.

Ainsi, pour chaque client et chaque catégorie préférée, le montant moyen par transaction a été estimé en inversant ce taux, puis perturbé à l'aide d'une loi Gamma afin de refléter la dispersion naturelle des paniers d'achat. Cette approche conserve une cohérence entre les montants dépensés et le cashback observé tout en générant une variabilité nécessaire à la robustesse de la modélisation.

5. Construction de la base transactionnelle

L'ensemble de ces étapes permet de générer une base transactionnelle détaillée, structurée sous la forme suivante :

- **CustomerID** : identifiant unique du client,
- **Date** : date estimée de la transaction,
- **Amount** : montant simulé du panier associé.

Cette base constitue l'entrée directe pour la fonction `summary_data_from_transaction_data` de la librairie *lifetimes*, qui calcule automatiquement les variables nécessaires aux modèles :

- **frequency** : nombre de transactions récurrentes (n−1n - 1n−1),
- **recency** : durée entre la première et la dernière transaction,
- **T** : durée entre la première transaction et la fin de la période d'observation,
- **monetary_value** : valeur moyenne des transactions du client.

6. Intégration de la durée de relation client

Enfin, pour garantir la cohérence entre les durées estimées par la reconstruction transactionnelle et les informations initiales, l'âge réel du client (*Tenure*) a été intégré et utilisé comme borne supérieure pour la variable T. Cela évite toute surestimation du temps observé et assure une meilleure représentativité des trajectoires clients.

En définitive, ce processus de préparation a permis de passer d'un jeu de données **agrégé et incomplet** à une base **transactionnelle reconstruite**, parfaitement adaptée à l'estimation du CLV via les modèles BG/NBD et Gamma-Gamma. Cette étape méthodologique, bien que fondée sur certaines hypothèses de simulation, représente un compromis rigoureux permettant d'exploiter au mieux les données disponibles et de garantir la comparabilité avec les approches standard de la littérature.

5.6.3 Application des modèles

Mise en place des modèles BG/NBD et Gamma-Gamma

L'estimation du Customer Lifetime Value (CLV) a été réalisée en combinant les modèles probabilistes **BG/NBD (Beta-Geometric/Negative Binomial Distribution)** pour la fréquence d'achat et **Gamma-Gamma** pour la valeur monétaire moyenne des transactions.

L'implémentation s'est appuyée sur la bibliothèque ***Lifetimes*** de Python, dédiée à la modélisation du comportement client, ainsi que sur les bibliothèques usuelles d'analyse et de visualisation telles que **Pandas**, **Matplotlib** et **Seaborn**.

Les étapes principales sont les suivantes :

- **Entraînement du modèle BG/NBD** sur les variables de fréquence, récence et durée d'observation.
- **Entraînement du modèle Gamma-Gamma** sur les clients dont la valeur monétaire moyenne est strictement positive.
- **Estimation du CLV sur 12 mois**, avec un taux d'actualisation fixé à 1% et une granularité journalière pour refléter la nature des données reconstruites.

Seuls **3087 clients sur les 5630 du dataset initial** ont été retenus pour l'estimation, conformément aux critères de préparation des données (ancienneté minimale et valeur monétaire positive).

Distribution du CLV sur 12 mois

L'estimation du CLV a généré une distribution relativement hétérogène, avec des disparités notables entre les clients.

Statistiques descriptives :

Indicateur	Valeur
Effectif (n)	3087
Moyenne	4432.75
Écart-type	1434.82
Minimum	0.46
25%	3389.86
Médiane (50%)	4061.02
75%	5332.17
Maximum	9279.77

Source : estimation CLV (BG/NBD + Gamma-Gamma), horizon 12 mois (Auteur)

Histogramme de la distribution du CLV :

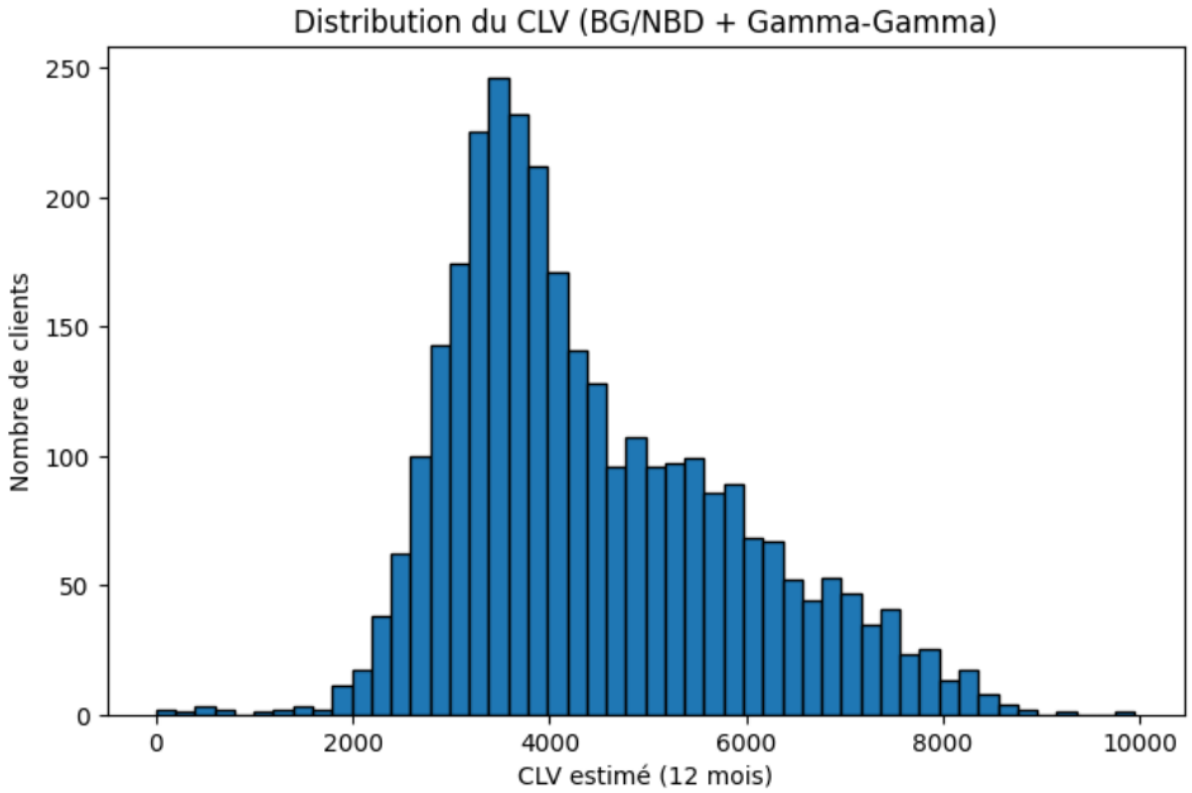


Figure X : Distribution du CLV (12 mois) estimé à l'aide des modèles BG/NBD + Gamma-Gamma

L'histogramme met en évidence une concentration des clients autour de la valeur médiane (≈ 4061), avec une dispersion importante vers des valeurs élevées (jusqu'à 9279).

Interprétation :

- La valeur médiane indique qu'un client « typique » génère environ **4061 unités monétaires par an**.
- La longue traîne vers la droite reflète l'existence d'une minorité de clients très rentables, constituant une cible prioritaire pour les stratégies marketing.
- Le caractère réaliste des résultats est renforcé par l'ordre de grandeur obtenu, bien que l'exactitude du modèle n'ait pu être vérifiée à l'aide de métriques telles que le **MSE** ou le **RMSE** (faute de données réelles de validation).

5.6.4 Segmentation des clients selon le CLV

Pour enrichir l'interprétation et faciliter la prise de décision, les clients ont été segmentés en **quartiles** de CLV, correspondant à quatre niveaux de valeur :

- **Faible valeur**
- **Valeur intermédiaire basse**
- **Valeur intermédiaire haute**
- **Haute valeur (VIP)**

Visualisation par boxplot :

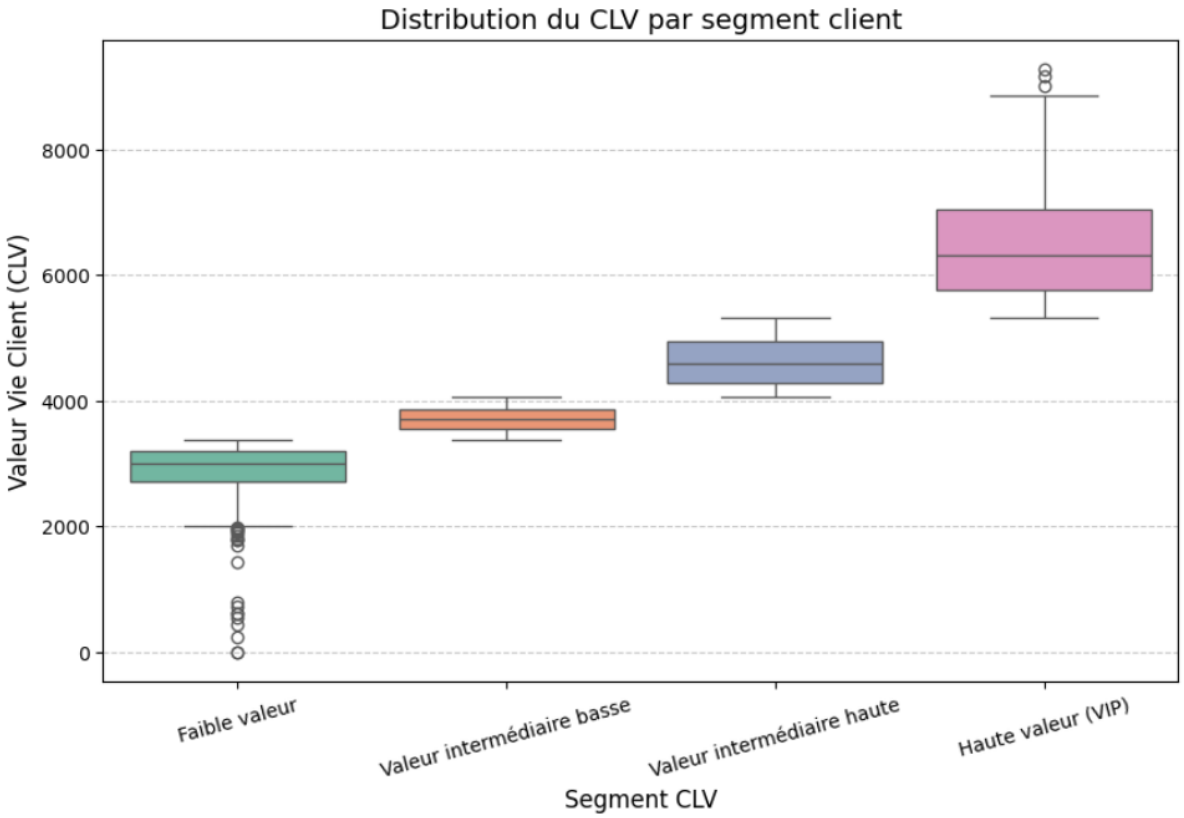


Figure X: Distribution du CLV par segment client – Boxplot

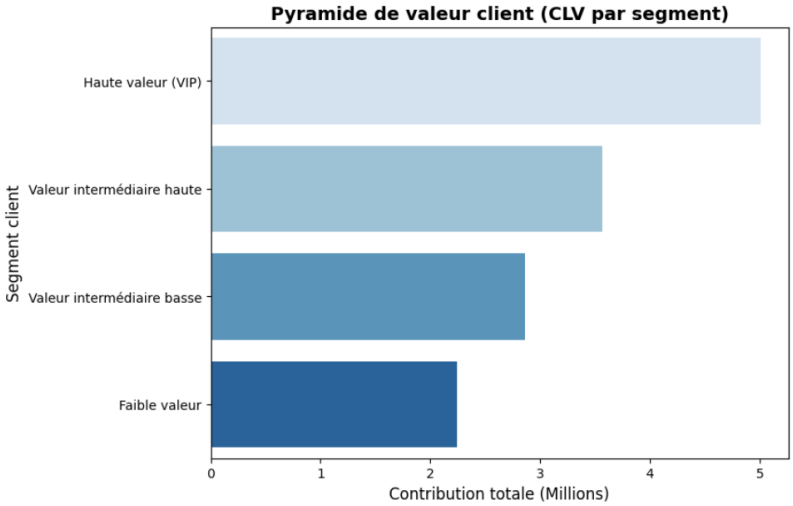
Le boxplot montre que :

- Le segment **VIP** concentre des valeurs largement supérieures à la médiane.
- Les segments intermédiaires présentent une répartition plus resserrée, mais contribuent de manière significative au chiffre d'affaires global.
- Le segment **faible valeur** reste majoritaire en effectif, mais contribue moins en termes de revenu.

Contribution des segments au chiffre d'affaires global

Afin de quantifier l'impact économique de chaque segment, la somme des CLV a été calculée par groupe.

Segment CLV	Contribution totale	Part relative
Haute valeur (VIP)	5,01 M	36,6 %
Valeur intermédiaire haute	3,57 M	26,1 %
Valeur intermédiaire basse	2,86 M	20,9 %
Faible valeur	2,24 M	16,4 %



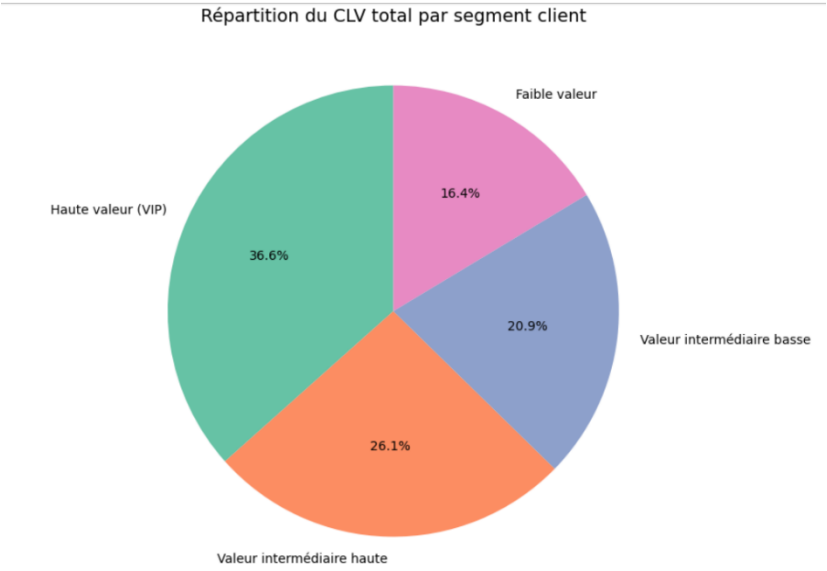


Figure X : Répartition du CLV total par segment client – Diagramme circulaire

Analyse :

- Les **clients VIP**, qui représentent seulement le quart supérieur de la base, génèrent **plus du tiers du revenu annuel**.
- Les segments intermédiaires représentent ensemble **près de 47 % du chiffre d'affaires**, confirmant leur importance stratégique.
- Le segment « faible valeur » contribue marginalement (16,4 %), suggérant que les investissements marketing ciblant cette population doivent être optimisés.

5.6.5 Analyse de concentration du revenu : courbe de Pareto

Afin d'évaluer la concentration de la valeur client, une **courbe cumulative de Pareto** a été construite à partir de la distribution estimée du CLV. Cette méthode permet de déterminer quelle proportion de clients contribue à une part donnée du chiffre d'affaires.

Les résultats montrent que :

- Pour générer **80 % du CLV total annuel**, il est nécessaire de mobiliser **2223 clients**, soit environ **39,5 % de la base client**.
- À l'inverse, les **60,5 % restants** ne contribuent qu'à 20 % du revenu.

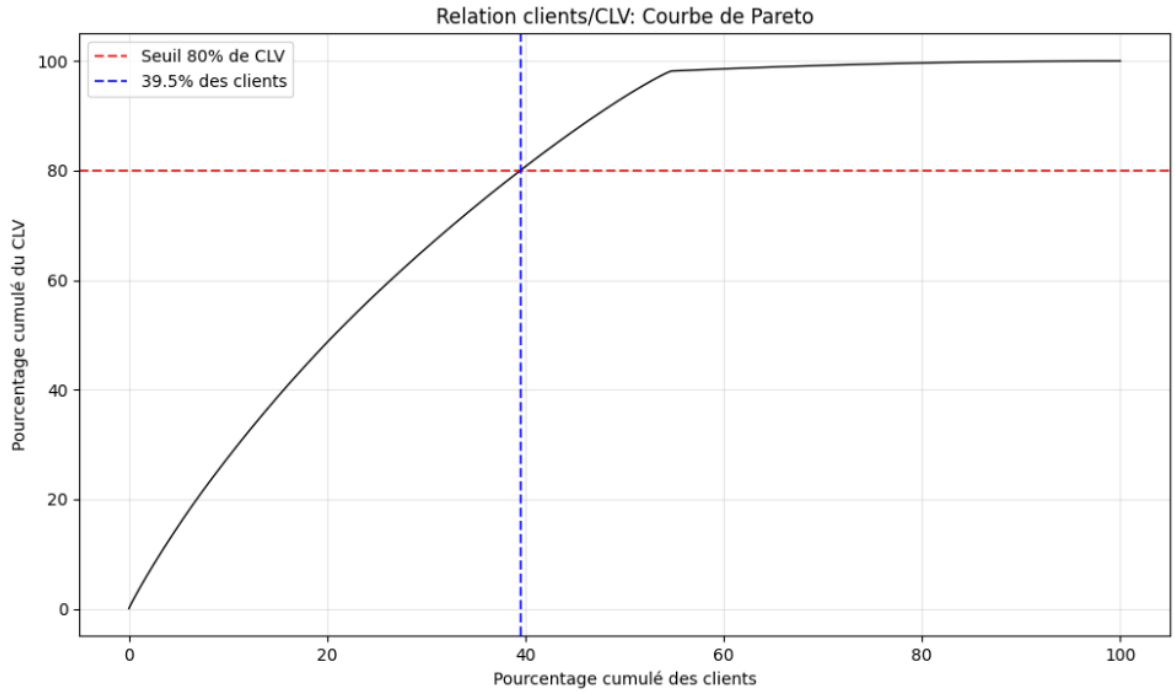


Figure X: Courbe cumulative de Pareto du CLV – Horizon 12 mois

Interprétation :

- Cette distribution confirme l'existence d'un phénomène de concentration économique classique : une minorité de clients génère la majorité du chiffre d'affaires.
- Bien que la répartition observée (≈ 40/60) soit moins extrême que la règle empirique des **20/80**, elle met néanmoins en évidence un groupe prioritaire représentant un **levier stratégique majeur** pour les politiques de fidélisation et de rétention.
- Ces clients à haute valeur (≈ deux cinquièmes de la base) devraient être ciblés par des actions marketing différenciées, telles que des programmes VIP, des offres personnalisées ou des avantages exclusifs, afin de sécuriser et maximiser leur contribution.

Conclusion

L'application conjointe des modèles **BG/NBD** et **Gamma-Gamma** a permis de fournir une estimation **réaliste et exploitable** de la valeur vie client (CLV) sur un horizon de 12 mois. Cette approche a révélé plusieurs enseignements majeurs :

- une **appréciation quantitative du CLV moyen**, offrant une vision globale de la rentabilité annuelle par client ;
- une **mise en évidence de l'hétérogénéité** du portefeuille, avec des disparités notables entre segments de faible valeur et clients à haute contribution ;
- l'identification de **segments stratégiques** (VIP et intermédiaire haut), constituant des cibles prioritaires pour la fidélisation et l'allocation différenciée des ressources marketing.

Par ailleurs, l'analyse de type **Pareto** a confirmé que la contribution au chiffre d'affaires est concentrée : environ **40 % des clients génèrent 80 % du CLV total**. Ce constat, bien qu'un peu moins marqué que la règle empirique 20/80, illustre clairement la nécessité pour l'entreprise de concentrer ses efforts sur la minorité de clients à forte valeur, tout en ajustant ses stratégies pour le reste du portefeuille.

Un autre apport essentiel de cette démarche réside dans la **capacité à estimer un CLV individuel pour chaque client**, permettant ainsi de les classer dynamiquement dans un segment donné et d'anticiper leur contribution potentielle. L'entreprise dispose donc d'un **outil décisionnel opérationnel** pour ajuster ses politiques marketing, concevoir des programmes de fidélité personnalisés et orienter les investissements en fonction de la rentabilité attendue.

Il convient toutefois de rappeler une **limite méthodologique** : faute de données réelles sur les montants dépensés et en l'absence de métriques de validation (telles que MSE ou RMSE), la fiabilité absolue des prédictions ne peut être vérifiée empiriquement dans ce cas. Néanmoins, selon la **littérature académique et les applications pratiques rapportées dans l'état de l'art**, la combinaison BG/NBD – Gamma-Gamma est largement reconnue comme l'une des méthodes les plus efficaces pour l'estimation du CLV. Dans notre contexte, elle permet donc de pallier les contraintes de données disponibles tout en offrant une **base solide de réflexion stratégique** sur la valeur client.

5.7 Synérgie des modèles

L'idée maîtresse de ce travail est de dépasser une logique de silos méthodologiques pour construire un **pipeline analytique intégré**, où chaque modèle contribue à enrichir et contextualiser les résultats des autres. En d'autres termes, il ne s'agit pas uniquement d'appliquer séparément des méthodes de segmentation, de classification, d'analyse de survie et d'estimation de la valeur vie client, mais de les faire interagir pour produire une information consolidée, directement exploitable dans la prise de décision stratégique.

Cette approche intégrative permet de transformer des analyses techniques complexes en leviers actionnables pour la stratégie d'entreprise.

L'architecture intégrative du pipeline analytique

Ainsi, la segmentation par **K-Means** permet de regrouper les clients selon leurs comportements, fournissant un premier niveau de lecture sur la structure de la base. Les modèles de classification, notamment **XGBoost**, apportent une mesure immédiate du risque de churn, identifiant avec une précision algorithmique les clients les plus susceptibles de quitter l'entreprise à court terme, tandis que l'analyse de survie (Kaplan-Meier, CoxPH, Weibull AFT) enrichit cette prédiction en introduisant une dimension temporelle, permettant d'anticiper la probabilité de départ dans un horizon de 3 ou 6 mois. Enfin, l'estimation de la **Customer Lifetime Value (CLV)** via le couple BG/NBD – Gamma-Gamma projette la valeur financière future attendue du client, consolidant la perspective économique.

L'approche unifiée permet de dépasser les limitations intrinsèques de chaque méthode :

- la segmentation seule ne quantifie pas le risque
- le scoring XGBoost seul ignore la dimension temporelle
- l'analyse de survie seule ne priorise pas par valeur économique
- le CLV seul n'identifie pas les clients à risque immédiat

La matrice décisionnelle unifiée : Du diagnostic à l'action

Cette intégration se matérialise par la capacité à générer un **jeu de données unifié** où chaque ligne correspond à un client, et synthétise les résultats des différents modules. Un exemple de sortie consolidée est donné ci-dessous :

CustomerID	segment	proba_xgboost	proba_churn_3mois	proba_churn_6mois	CLV_prédit_12mois	churn_p
50001	1	0.938886	0.169527	0.217408	1682.002257	1
50002	1	0.984727	0.527916	0.628630	279.476261	1
50003	1	0.990481	0.738430	0.829629	594.035246	1
50004	1	0.996642	0.199459	0.254416	1126.120314	1
50005	1	0.954968	0.216940	0.275827	1411.741313	1

Ce tableau illustre concrètement la valeur ajoutée de la synergie des modèles :

- chaque client est **rattaché à un segment comportemental** (K-Means),
- il dispose d'une **probabilité instantanée de churn** estimée par XGBoost,
- des **probabilités conditionnelles de départ** dans 3 et 6 mois (Weibull AFT),
- et enfin une **estimation de sa valeur vie à 12 mois (CLV)**, indicateur clé pour hiérarchiser les actions de rétention.

Au-delà des interprétations spécifiques déjà discutées pour chaque méthode individuellement (ex. : lecture des courbes Kaplan-Meier, identification des covariables significatives par CoxPH), cette **vision unifiée** offre à l'entreprise un instrument stratégique puissant. Elle permet notamment :

- d’identifier non seulement les clients à risque, mais aussi **ceux dont la perte engendrerait la plus forte valeur détruite** ;
- de **prioriser les interventions marketing** (ex. : cibler en premier les clients à forte CLV avec un risque de churn élevé) ;
- de concevoir des **scénarios de rétention différenciés** selon les segments détectés ;
- et de disposer d’un outil évolutif, facilement intégrable dans des systèmes décisionnels ou des tableaux de bord opérationnels.

En définitive, cette synergie représente l’apport majeur du projet : la **traduction d’une approche scientifique rigoureuse en un dispositif opérationnel concret**, rare dans le contexte des entreprises où les études restent souvent limitées à des modèles isolés. Ce pipeline unifié ouvre ainsi la voie à une exploitation plus fine des données clients, conciliant **performance algorithmique, pertinence économique et applicabilité pratique**.

Chapitre 6 – Recommandations stratégiques et perspectives

Introduction

L’analyse du churn n’acquiert sa pleine pertinence que lorsqu’elle débouche sur des actions concrètes, ciblées et mesurables. L’objectif de ce chapitre est de traduire les résultats issus des différentes approches analytiques en recommandations stratégiques et en leviers opérationnels adaptés au contexte de l’entreprise. En mobilisant la complémentarité de ces méthodes, il s’agit de proposer un cadre méthodologique intégrant à la fois l’identification des clients à risque de départ, la priorisation en fonction de leur valeur financière, ainsi que la temporalité associée à ce risque.

Plutôt qu’une approche uniforme et générique, les recommandations seront différenciées en fonction des segments de clientèle, de manière à garantir que chaque action de fidélisation corresponde au profil et aux comportements spécifiques des clients concernés. L’articulation entre la probabilité de churn estimée par XGBoost, l’horizon temporel fourni par l’analyse de survie et la dimension économique apportée par la CLV permettra d’optimiser l’allocation des ressources et de maximiser l’impact des interventions.

Ce chapitre intégrera par ailleurs des visualisations explicites afin de mettre en évidence la contribution de chaque segment à la valeur monétaire globale ainsi qu'au risque de churn. Il se conclura par des recommandations stratégiques et des perspectives de recherche et de mise en œuvre, visant à inscrire durablement l'analyse prédictive comme un axe central de la stratégie de croissance et de fidélisation de l'entreprise.

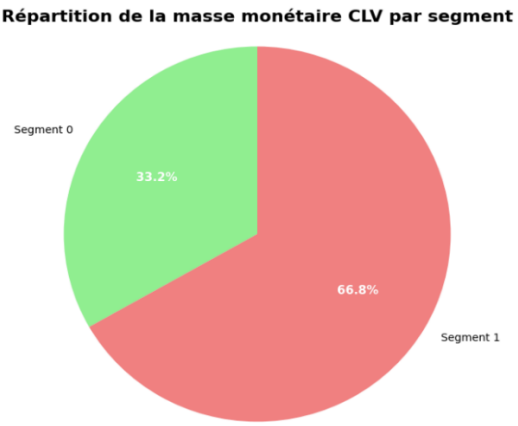
6.1 Recommandations stratégiques fondée sur le segment

La segmentation réalisée par **K-Means (k = 2)** a permis de distinguer deux grandes catégories de clients, intégrées comme variable explicative dans les modèles de classification :

- **Segment 0** : clientèle relativement stable, présentant un faible risque de churn et une survie moyenne plus élevée.
- **Segment 1** : clientèle instable, caractérisée par une probabilité accrue de churn et une survie médiane plus courte.

Au-delà des dynamiques de rétention, l'analyse de la **masse monétaire** associée à chaque segment met en évidence un déséquilibre marqué dans la distribution de la valeur :

- **Segment 0** : 8,1 M unités monétaires, soit **33,2 %** du total.
- **Segment 1** : 16,3 M unités monétaires, soit **66,8 %** du total.
- **Masse monétaire totale** : 24,5 M unités monétaires



Cette répartition révèle que, malgré leur instabilité et leur risque élevé de churn, les clients du **Segment 1 concentrent plus des deux tiers de la valeur financière totale générée**. En d'autres termes, la base de revenus de l'entreprise repose majoritairement sur une clientèle fragile et susceptible de quitter prématurément. À l'inverse, le Segment 0, plus stable, génère une part minoritaire de la valeur, bien qu'il constitue une zone de réassurance.

6.1.2 Cartographie churn × valeur vie client (CLV)

L'un des apports majeurs de ce travail est le pouvoir de mettre en place d'un quadrillage stratégique, par exemple à **4 cases opérationnelles** :

- **Haut CLV / Haut risque** : clients **prioritaires** qui offres de rétention personnalisées, gestion proactive via agents dédiés.
- **Haut CLV / Bas risque** : clients **premium**, programmes VIP, avantages exclusifs, communication valorisante.
- **Bas CLV / Haut risque** → clients secondaires : actions automatisées et peu coûteuses (emails, notifications, bons de réduction ciblés).
- **Bas CLV / Bas risque** → clients peu stratégiques : surveillance minimale, maintien automatisé avec optimisation des coûts.

Cette matrice permet d'**allouer les ressources marketing de façon rationnelle**, en maximisant le retour sur investissement des campagnes de fidélisation.

6.1.3 Typologies de profils critiques (exemples basés sur l'analyse)

Les analyses menées (classification et survie) permettent d'identifier plusieurs profils de clients critiques, qu'il convient de cibler en priorité :

- **Clients plaignants** : la variable *Complain* est un facteur fortement associé au churn (CoxPH : hazard ratio ↑ ; Weibull AFT : réduction significative du temps de survie).
- **Clients du Segment 1** : Kaplan–Meier montre une chute rapide de la survie, sous 0,80 dès le 10e mois.
- **Préférence de paiement COD (Cash on Delivery)** : mode de paiement corrélé à un risque accru de départ, probablement en lien avec une faible fidélisation.
- **Dispositif de connexion** : certaines modalités, comme *PreferredLoginDevice_Mobile Phone*, apparaissent protectrices (HR < 1), révélant un usage plus ancré et régulier.

Recommandation : établir des **"profils d'alerte"** sous forme de règles métier, combinant les dimensions suivantes :

- score de churn (XGBoost) ;
- valeur vie prédite (CLV) ;
- variables explicatives critiques (plaintes, mode de paiement, ancienneté depuis la dernière commande).

Ces profils permettront à l'entreprise d'**industrialiser les actions de fidélisation**, en déployant des campagnes ciblées et différenciées selon la typologie des clients.

6.2 Priorisation des actions de fidélisation

6.2.1 Cartographie churn × valeur vie client (CLV)

Code : 0000000000TravfinalPFEcOMPLET.ipynb + 5.7 synergie

6.3 Limites

Doc : rapport PFE Master

Limites : clv non vérifiés, manque de données

6.4 Perspectives et pistes d'amélioration

Conclusion générale

Ce mémoire a proposé une démarche intégrée et rigoureuse pour traiter la problématique de la prédiction et de la gestion du churn, en articulant de manière complémentaire plusieurs approches quantitatives issues de la **science des données**, de la **modélisation mathématique** et du **machine learning**.
Dans un premier temps, la **classification supervisée** a permis d'identifier avec une grande fiabilité les clients les plus exposés au risque de churn. Le modèle retenu, XGBoost, s'est distingué par ses performances prédictives élevées et sa capacité à fournir un outil de scoring directement mobilisable dans un contexte opérationnel. Cette première contribution ancre le travail dans une logique de décision pragmatique et immédiate pour l'entreprise.

Au-delà de l'identification des clients « à risque », **l'analyse de survie** a enrichi la compréhension de la dynamique temporelle du churn. Les méthodes employées :Kaplan-Meier, modèles semi-paramétriques de Cox et modèles paramétriques de type Weibull AFT ont permis d'estimer la probabilité de rétention en fonction du temps, tout en mettant en évidence les facteurs qui accélèrent ou ralentissent le départ des clients. Ces résultats offrent une lecture nuancée de la problématique : non seulement savoir qui est susceptible de partir, mais surtout à quel moment et sous quelles conditions.

Enfin, l'intégration des modèles BG/NBD et Gamma-Gamma a permis de dépasser la seule logique du risque pour **quantifier la valeur économique attendue** de chaque client à l'horizon futur. L'estimation du Customer Lifetime Value (CLV) a mis en évidence des disparités majeures dans la contribution des segments de clientèle, comme l'illustre la courbe de Pareto où environ 39,5 % des clients génèrent 80 % de la valeur totale. Ce résultat apporte un cadre objectif et robuste pour la priorisation budgétaire et la définition de stratégies de rétention différenciées.

La combinaison de ces trois approches (identification, temporalité et valorisation) constitue la principale originalité et valeur ajoutée de ce travail. Elle offre à l'entreprise un dispositif décisionnel complet : savoir quels clients cibler, quand agir et combien investir. Sur le plan pratique, cette articulation se traduit par des recommandations stratégiques opérationnalisables : segmentation conjointe churn × CLV, mise en place de plans d'action différenciés, élaboration de tableaux de bord dynamiques et définition d'indicateurs de suivi pertinents.

Certes, certaines limites doivent être reconnues, liées notamment à la nature des données disponibles (agrégations, proxys pour les montants) et aux hypothèses structurelles des modèles retenus. Toutefois, la méthodologie adoptée, caractérisée par la robustesse des cadres théoriques, la confrontation de plusieurs approches et la validation systématique des résultats garantit la solidité scientifique des conclusions et leur pertinence opérationnelle.

En définitive, ce mémoire apporte une contribution double : d'un côté, il enrichit la littérature sur l'application combinée de modèles de machine learning, d'analyse de survie et de modélisation probabiliste en gestion de la relation client ; de l'autre, il fournit à l'entreprise un socle méthodologique et stratégique susceptible de transformer les résultats analytiques en gains mesurables. Sa mise en œuvre progressive, via des expérimentations pilotes et un suivi rigoureux, ouvre la voie à une industrialisation durable de la gestion du churn et de la valorisation client, offrant ainsi des perspectives prometteuses tant pour le pilotage opérationnel que pour la recherche future.

REFERENCES BIBLIOGRAPHIQUES ET WEBOGRAPHIQUES

[1] K. Khadka and S. Maharjan, ‘‘Customer satisfaction and customer loyalty,’’ Centria Univ. Appl. Sci. Pietarsaari, vol. 1, no. 10, pp. 58–64, 2017.

[2] J. Ahn, J. Hwang, D. Kim, H. Choi, and S. Kang, ‘‘A survey on churn analysis in various Bus. Domains,’’ IEEE Access, vol. 8, pp. 220816–220839, 2020.

[3] J.N. Shethand C. Uslay,‘‘Creatingenduringcustomervalue,’’J.Creating Value, vol. 8, no. 2, pp. 241–252, Nov. 2022.

[4] T. Dierkes, M. Bichler, and R. Krishnan, ‘‘Estimating the effect of word of mouth on churn and cross-buying in the mobile phone market with Markov logic networks,’’ Decis. Support Syst., vol. 51, no. 3, pp. 361–371, Jun. 2011.

[5] K. Ljubičić, A. Merćep, and Z. Kostanjčar, ‘‘Churn prediction methods based on mutual customer interdependence,’’ J. Comput. Sci., vol. 67, Mar. 2023, Art. no. 101940.

[6] Reichheld, F.F.; Sasser, W.E. Zero defeofions: Quoliiy comes to services. Harvard. Bus. Rev. 1990, 68, 105–111. Jones, T.O.; Sasser, W.E., Jr. Why satisfied customers defect. IEEE Eng. Manag. Rev. 1998, 26, 16–26.

[76] T. Gattermann-Itschert and U. W. Thonemann, ‘‘Proactive customer retention management in a non-contractual B2B setting based on churn prediction with random forests,’’ Ind. Marketing Manage., vol. 107, pp. 134–147, Nov. 2022.

[77] M. Mirkovic, T. Lolic, D. Stefanovic, A. Anderla, and D. Gracanin, ‘‘Customer churn prediction in B2B non-contractual Bus. Settings using invoice data,’’ Appl. Sci., vol. 12, no. 10, p. 5001, May 2022.

[91] S. Tavassoli and H. Koosha, ‘‘Hybrid ensemble learning approaches to customer churn prediction,’’ Kybernetes, vol. 51, no. 3, pp. 1062–1088, Feb. 2022.

[105] M. A. S. Thorat and V. R. Sonawane, ‘‘A random forest churn prediction model: An study of machine learning techniques for churn prediction and factor identification in the telecom sector,’’ Scandin. J. Inf. Syst., vol. 35, no. 1, pp. 818–824, 2023.

[113] A. M A and S. K K, ‘‘An efficient hybrid classifier model for customer churn prediction,’’ Int. J. Electron. Telecommun., vol. 69, pp. 11–18, Dec. 2022.

[118] M. Milošević, N. Živić, and I. Andjelković, ‘‘Early churn prediction with personalized targeting in mobile social games,’’ Expert Syst. Appl., vol. 83, pp. 326–332, Oct. 2017.

[120] R. A. de Lima Lemos, T. C. Silva, and B. M. Tabak, ‘‘Propension to customer churn in a financial institution: A machine learning approach,’’ Neural Comput. Appl., vol. 34, no. 14, pp. 11751–11768, Jul. 2022.

[125] S. M. Sina Mirabdolbaghi and B. Amiri, ‘‘Model optimization analysis of customer churn prediction using machine learning algorithms with focus on feature reductions,’’ Discrete Dyn. Nature Soc., vol. 2022, pp. 1–20, Jun. 2022.

[126] R. Wirth and J. Hipp, ‘‘Crisp-DM: Towards a standard process model for data mining,’’ in Proc. 4th Int. Conf. practical Appl. Knowl. Discovery Data Mining, vol. 1, 2000, pp. 29–39.

[143] D. AL-Najjar, N. Al-Rousan, and H. AL-Najjar, ‘‘Machine learning to develop credit card customer churn prediction,’’ J. Theor. Appl. Electron. Commerce Res., vol. 17, no. 4, pp. 1529–1542, Nov. 2022.

[168] Awais Manzoor, M. Atif Qureshi, Etain Kidney, Luca Longo, “A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners”, May 2024

[170] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, ‘‘A comparison of machine learning techniques for customer churn prediction,’’ Simul. Model. Pract. Theory, vol. 55, pp. 1–9, Jun. 2015.

[171] E. M. Elgohary, M. Galal, A. Mosa, and G. A. Elshabrawy, ‘‘Smart evaluation for deep learning model: Churn prediction as a product case study,’’ Bull.

[200] M. Alizadeh, D. S. Zadeh, B. Moshiri, and A. Montazeri, ‘‘Development of a customer churn model for banking industry based on hard and soft data fusion,’’ IEEE Access, vol. 11, pp. 29759–29768, 2023.

Chapiter 3:
[3-4-X1] N. Lu, H. Lin, J. Lu, and G. Zhang, "A customer churn prediction modelin telecom industry using boosting," IEEE Transactions on Industrial Informatics, vol. 10, no. 2, pp. 1659–1665, 2012.

[3-4-X2] B. Huang, M. T. Kechadi, and B. Buckley, "Customer churn predictionin telecommunications," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1414–1425, 2012.

[3-4-3] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972
[3-4-26] Anny Xiang, Pablo Lapuerta, Alex Ryutov, Jonathan Buckley, and Stanley Azen. Comparison of the performance of neural network methods and cox regression for censored survival data. Computational statistics & data analysis, 34(2):243–257, 2000

[3-4-4] David Faraggi and Richard Simon. A neural network model for survival data. Statistics in medicine, 14(1):73–82, 1995

[3-4-15] Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neuralnetworks and cox regression. Journal of machine learning research, 20(129):1–30, 2019

[3-4-8] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. 2008

[3-4-17] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018
[3-4-9] *Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. BMC medical research methodology, 18:1–12, 2018.*

[3-4-24] *Wafaa Tizi and Abdelaziz Berrado. Machine learning for survival analysis in cancer research: A comparative study. Scientific African, 21:e01880, 2023.*

[3-4-26] *Anny Xiang, Pablo Lapuerta, Alex Ryutov, Jonathan Buckley, and Stanley Azen. Comparison of the performance of neural network methods and cox regression for censored survival data. Computational statistics & data analysis, 34(2):243–257, 2000.*
[3-4-28] *Xulin Yang, Hang Qiu, Liya Wang, and Xiaodong Wang. Predicting colorectal cancer survival using time-to-event machine learning: Retrospective cohort study. Journal of Medical Internet Research, 25:e44417, 2023*

Chapiter 4
[15-4] Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. Journal of machine learning research, 20(129):1–30, 2019.
Chapiter 3
[1-4] Chen & Guestrin (KDD 2016), Bentéjac et al. (AI Review 2021) 2021
[2-4] Zhang et al. (IEEE Access 2022), Kaggle State of ML (2023
[3-4] Lundberg & Lee (NeurIPS 2017), Molnar (Interpretable ML 2022)
[4-4] Chen, 2016 ; Probst et al., 2019
[5-4] NVIDIA MLPerf Benchmarks (2023)
[19-4] C. Davidson-Pilon, "lifelines: survival analysis in python," Journal of Open Source Software, vol. 4, no. 40, p. 1317, 2019.
[25 -4] M. Zhou, Empirical Likelihood Method in Survival Analysis. Chapman & Hall/CRC Biostatistics Series, CRC Press, 2015.
[26-4] R. C. Paul Meier, Theodore Karrison and H. Xie, "The price of kaplan–meier," Journal of the American Statistical

Association, vol. 99, no. 467, pp. 890–896, 2004.
[27-4] **Kalbfleisch, J. D., & Prentice, R. L. (2002).** *The Statistical Analysis of Failure Time Data* (2nd ed.). Wiley.
[28-4]**Collett, D. (2015).** *Modelling Survival Data in Medical Research* (3rd ed.). Chapman & Hall/CRC.
[29-4]**Klein, J. P., & Moeschberger, M. L. (2006).** *Survival Analysis: Techniques for Censored and Truncated Data* (2nd ed.). Springer.
[30-4]**Wei, L. J. (1992).** The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*, 11(14-15), 1871–1879.
[31-4]**Lawless, J. F. (2003).** *Statistical Models and Methods for Lifetime Data* (2nd ed.). Wiley.
[32-4]**Bender, R., Augustin, T., & Blettner, M. (2005).** Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24(11), 1713–1723.
[33-4]**Cox, D. R., & Oakes, D. (1984).** *Analysis of Survival Data*. Chapman & Hall.
[35-4]**lifelines (Python)** – Documentation de WeibullAFTFitter : implémentation pratique moderne des modèles AFT avec régression potentielle sur les paramètres d'échelle et de forme.
[36-4] Zhang (2016) – Parametric regression model for survival data
[37-4] Amico and Van Keilegom (2014) – Survival analysis and regression models
[38-4] Yigzaw Alemu Limenih & Demeke Lakew Workie (2019) – Survival analysis of time to cure on multidrug resistance tuberculosis patients in Amhara region, Ethiopia
[39-4] Zhang (2016) – Parametric regression model for survival data: Weibull regression model as an example
[4-12X] H. Casteran, L. Mayer-Waarden and W. Reinartz "Modeling Customer Lifetime Value, Retention, and Churn". In: Homburg C., Klarmann M., Vomberg A. (eds),Handbook of Market Research. Springer, Cham. p.14-41, April 2017.
[4-13X] M. Mzoughia and M. Limam, "An Improved BG/NBD, Approach for Modeling Purchasing Behavior Using ComPoisson Distribution" *Internation Journal of Modeling and Optimization*, vol 4. no. 2. pp. 141-145, 2014 [chap 4]
Références bibliographiques POUR EVALUATION BRIER SCORE ET CONCORDANCE

- Brier, G. W. (1950). *Verification of forecasts expressed in terms of probability*. Monthly Weather Review, 78(1), 1-3.
- Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). *Assessment and comparison of prognostic classification schemes for survival data*. Statistics in Medicine, 18(17-18), 2529–2545.
- Gerds, T. A., & Schumacher, M. (2006). *Consistent estimation of the expected Brier score in general survival models with right-censored event times*. Biometrical Journal, 48(6), 1029–1040.
- Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). *Evaluating the yield of medical tests*. JAMA, 247(18), 2543–2546.

[1-5] Provost & Fawcett (2013). Data Science for Business. O'Reilly.
[2-5] Neslin et al. (2006). "Defensive Marketing Strategy Using Customer Churn Prediction". Marketing Science.
[3-5] Rust & Huang (2014). "The Service Revolution and the Transformation of Marketing Science". Journal of Marketing Research.
[4-5] Fader & Hardie (2009). "Probability Models for Customer-Base Analysis". Journal of Interactive Marketing.
[4-X1] Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2004). "Counting Your Customers" the Easy Way: An Alternative to the Pareto/NBD Model. Marketing Science, Working Paper.
[4-X2] Schmittlein, D. C., Morrison, D. G., & Colombo, R. (1987). Counting Your Customers: Who They Are and What Will They Do Next? Management Science, 33(1), 1-24 (chapitre 4)
[4-10X] P. Fader, B. Hardie and K. Lee. "Counting YourCustomers, the Easy Way: An Alternative to the Pareto/NBD Model". *Marketing Science,* vol. 24, no.2, pp. 275-284, 2005.
[5-5] Verbraken et al. (2013). "Profit-Based Model Selection for Customer Churn Prediction". Expert Systems with Applications.
[6-5] Kuhn & Johnson (2013) : "Applied Predictive Modeling" (Springer), p. 42
[7-5] Kuhn & Johnson (2013) : "Applied Predictive Modeling" (Springer), p. 84
[8-5] Zheng & Casari (2018) : "Feature Engineering for Machine Learning" (O'Reilly), chap 4
[9-5] Harrison & Rubinfeld (1978) : *"Hedonic Housing Prices and the Demand for Clean Air"* (Journal of Environmental Economics)
[10-5] Breiman (2001) : "Random Forests" (Machine Learning Journal)
[11-5] Fernández et al. (2018) : "Learning from Imbalanced Data Sets" (Springer), chap. 5
[12-5] Chawla et al. (2002) : "SMOTE: Synthetic Minority Over-sampling Technique" (JAIR)
[13-5] Kohavi (1995) : "A Study of Cross-Validation and Bootstrap for Accuracy Estimation"
[36-chap5-4.] Wu, J.; Shi, L.; Yang, L.P.; Niu, X.X.; Li, Y.Y.; Cui, X.D.; Tsai, S.B.; Zhang, Y.B. User Value Identification Based on Improved RFM Model and K-Means++ Algorithm for Complex Data Analysis. *Wirel Commun. Mob.Com.* **2021**, *9982484*, 1–8
[37-chap5-4.] Li, Y.; Chu, X.Q.; Tian, D.; Feng, J.Y.; Mu, W.S. Customer segmentation using K-means clustering and the adaptive. Appl. Soft Comput. 2021, 113, 107924.
[40-chap5-4.] Hosseini, M.; Shajari, S.; Akbarabadi, M. Identifying multi-channel value co-creator groups in the banking industry. J. Retail. Consum. Serv. 2022, 5, 102312.
[41-chap5-4.] Zhou, J.; Zhai, L.L.; Pantelous, A.A. Market Segmentation Using High-dimensional Sparse Consumers Data. *Expert. Syst. Appl.* **2020**, *145*, 113136.