**Dakar Institute of Technology**

****

**Habilitation n°3386-21 FEV.2024**

**Domaine : Sciences et Technologies**

**Département : Informatique**

**Spécialité : Intelligence Artificielle**

**Projet Informatique**

Présenté par

**………………………………………**

**SUJET :**

**Soutenu à……… le .../.../...**

**Année Académique : 20//-20//**

# DÉDICACE

À la mémoire de mon père, **KPAVODE Athanase**, Ton absence m’a appris la force, ton exemple continue de m’inspirer. Que ce travail soit un témoignage silencieux de ton héritage dans ma vie.

À ma mère, **KPODAGBAME Eugénie**, Pour ton amour constant, ta patience et ta lumière dans mes moments de doute. Tu es mon pilier et ma source inépuisable de courage.

À mes frères et sœur, **Gérald**, **Osvalt**, **Morel** et **Mariselle**, Merci pour votre affection, votre présence et vos encouragements précieux. Ce mémoire est aussi le fruit de notre histoire commune et de votre soutien discret.

À mes enseignants, Pour m’avoir transmis bien plus que des connaissances : une rigueur intellectuelle, une passion pour la recherche, et une exigence noble envers moi-même. Votre accompagnement a été déterminant dans l’aboutissement de ce travail.

À mes collègues de la **BOA Bénin**, Pour votre compréhension, vos conseils, et votre bienveillance tout au long de ce parcours. Travailler à vos côtés m’a enrichi humainement et professionnellement.

À tous ceux qui, de près ou de loin, ont contribué à cette aventure académique, Cette œuvre vous est dédiée avec gratitude, humilité, et affection profonde.

# REMERCIEMENTS

Ce mémoire marque l’aboutissement d’un parcours académique et personnel nourri de rencontres, d’efforts et de soutien constant. Je tiens à exprimer ici ma gratitude à toutes celles et ceux qui ont contribué à sa réalisation.

À **Dieu Tout-Puissant**, pour la force, l’inspiration et les opportunités qui m’ont permis d’aller jusqu’au bout.

À **ma famille**, en particulier ma mère **KPODAGBAME Eugénie**, mes frères **Gérald**, **Osvalt**, **Morel** et ma sœur **Mariselle**, pour leur présence affective, leurs encouragements et leur foi en moi. À la mémoire de mon père, **KPAVODE Athanase**, dont l’héritage continue de guider mes pas.

À **l’ensemble de mes enseignants** et encadrants, pour la rigueur intellectuelle, les connaissances transmises et le suivi pédagogique qui ont façonné ce travail. Vos conseils et votre disponibilité ont été précieux.

À **mes collègues de la BOA Bénin**, pour leur bienveillance, leur patience et leurs encouragements tout au long de ce processus. Merci pour votre compréhension et votre accompagnement discret mais déterminant.

À **mes camarades de promotion et amis**, pour les échanges enrichissants, les révisions partagées et les moments de solidarité qui ont rythmé ce parcours.

Enfin, à toutes celles et ceux qui, de près ou de loin, ont participé à cette aventure académique, je vous adresse mes plus sincères remerciements.

# GLOSSAIRE

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# LISTE DES FIGURES

# 

# 

# 

# 

# 

# 

# LISTE DES TABLEAUX

# 

# 

# RÉSUMÉ

Ce mémoire s’inscrit dans une démarche analytique visant à anticiper le **comportement de résiliation client (churn)** dans le secteur bancaire, en exploitant les leviers offerts par l’**intelligence artificielle**. À travers une approche de modélisation supervisée, l’étude s’est appuyée sur un jeu de données client pour construire des outils prédictifs performants et interprétables.

L’étude a suivi plusieurs étapes clés :

* Une **analyse exploratoire des données** pour identifier les variables influentes dans le départ des clients.
* La mise en œuvre de **modèles de classification** tels que la **régression logistique**, le **Random Forest** et le **CatBoost**.
* L’utilisation de techniques de **rééquilibrage**, de **validation croisée**, et d’**interprétabilité (SHAP)** pour affiner les résultats.
* Le développement d’une **application interactive (Streamlit)** permettant une utilisation opérationnelle du modèle prédictif.

Les performances obtenues sont remarquables, avec un **AUC supérieur à 0.998** pour le Random Forest, confirmant sa capacité à discriminer les clients fidèles des churners.

Ce travail ouvre la voie à plusieurs **perspectives d’amélioration**, notamment l’intégration d’informations textuelles, l’analyse séquentielle des comportements, ou le déploiement du modèle en production via API sécurisée.

# ABSTRACT

# 

In a rapidly evolving banking landscape, retaining customers has become a strategic priority. This study presents an intelligent and interpretable machine learning solution designed to predict customer churn with high precision. Using supervised classification models—including Logistic Regression, Random Forest, and CatBoost—the research explores the predictive capacity of structured customer data.

The methodology involved exploratory analysis to identify impactful features, followed by model training with appropriate class balancing and validation strategies. Among the models tested, Random Forest emerged as the top performer, achieving an exceptional **Area Under the Curve (AUC) score exceeding 0.998**, highlighting its discriminative power.

To enhance transparency, model interpretation was facilitated through SHAP value analysis and feature importance plots. The integration of these insights into a **Streamlit-based application** allowed practical deployment, enabling real-time scoring and decision support.

This work lays the groundwork for future applications in real-time customer retention systems and opens perspectives for further improvements, including text data integration, sequential modeling, and cloud deployment via secure APIs.

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# SOMMAIRE

# 

# 

# 

# 

# 

# 

# 

# 

# INTRODUCTION

### CONTEXTE ET JUSTIFICATION

Dans un environnement bancaire marqué par une intensification de la concurrence et une exigence croissante des clients, la fidélisation devient un levier stratégique crucial pour les institutions financières. La qualité de service perçue par les clients constitue aujourd’hui un facteur déterminant dans leur satisfaction, leur engagement et leur rétention.

Cependant, dans le contexte actuel de notre banque, une augmentation significative des demandes de clôture de comptes a été observée, révélant un niveau de mécontentement préoccupant. Ce phénomène met en lumière la nécessité d’identifier et de comprendre les causes profondes de cette attrition, afin de mettre en place des mécanismes de prévention adaptés.

Les avancées récentes en matière d’analyse de données et d’intelligence artificielle offrent des opportunités inédites pour appréhender ces problématiques. Elles permettent notamment de détecter les signaux faibles de désengagement et d’élaborer des stratégies de fidélisation personnalisées et proactives.

### OBJECTIFS DE LA RECHERCHE

Ce mémoire vise à étudier les déterminants de l’attrition des clients au sein d’une banque à travers une approche analytique et prédictive. L’objectif principal est de construire un modèle de machine learning capable d’identifier les clients les plus susceptibles de clôturer leur compte, en exploitant un jeu de données issu de la plateforme Kaggle.

De manière plus spécifique, il s’agit de :

* Identifier les variables explicatives les plus influentes dans la décision de quitter la banque ;
* Mettre en œuvre un algorithme de classification performant pour prédire l’attrition ;
* Proposer des recommandations concrètes pour améliorer la qualité de service et réduire le taux de départs.

### STRUCTURE DU PROJET INFORMATIQUE

Le présent mémoire est structuré en quatre grandes parties :

* Un état de l’art présentant les travaux existants sur l’attrition client, les approches analytiques et les outils de machine learning utilisés dans ce contexte ;
* Une section méthodologique décrivant les étapes de traitement et d’analyse des données ;
* Une phase d’implémentation du modèle prédictif, avec une évaluation de ses performances ;
* Une dernière partie dédiée aux recommandations pratiques à destination de la banque, en lien avec les résultats obtenus.

### CONCLUSION

À travers cette démarche, ce projet entend allier rigueur académique et applicabilité terrain, en apportant des solutions innovantes et fondées sur les données pour renforcer la fidélité client. Il s’inscrit dans une volonté de transformation digitale des pratiques décisionnelles dans le secteur bancaire.

CHAPITRE I: Etat de l’art

## État de l'art sur le domaine d'application

Le secteur bancaire est en constante évolution, influencé par la digitalisation accélérée des services financiers, l'émergence de nouveaux acteurs disruptifs (fintech, néo-banques, challenger banks) et l'évolution rapide des comportements et attentes des clients, particulièrement depuis la pandémie de COVID-19. Dans ce contexte hautement concurrentiel et en perpétuelle mutation, la fidélisation des clients devient un enjeu crucial et stratégique. La perte d'un client représente non seulement une perte financière directe substantielle (estimée en moyenne entre 3 à 10 fois le coût d'acquisition selon les études du secteur), mais également un risque d'atteinte à la réputation de l'institution, surtout à l'ère des réseaux sociaux et de la transparence numérique où l'expérience client négative peut se propager instantanément et massivement.

L'attrition, ou churn, désigne le phénomène par lequel un client met fin à sa relation commerciale avec une entreprise. Dans le domaine bancaire, cela se traduit généralement par la clôture du ou des comptes détenus, le transfert des fonds vers un concurrent, ou l'abandon progressif des services. Les causes de l'attrition sont multiples et interconnectées : insatisfaction vis-à-vis du service client ou de la qualité de l'accompagnement, frais bancaires jugés trop élevés ou manque de transparence tarifaire, offres concurrentes plus attractives en termes de taux ou de services digitaux, absence de relation personnalisée et de conseils adaptés aux besoins spécifiques, incidents bancaires répétés (pannes de système, erreurs de traitement), ou encore inadéquation entre l'évolution des besoins du client et l'offre proposée par la banque.

Afin de mieux comprendre et prévenir ce phénomène complexe, les banques ont progressivement recours aux technologies avancées de traitement de données et d'intelligence artificielle. L'analyse prédictive appliquée à l'attrition permet d'anticiper les départs potentiels en s'appuyant sur l'exploitation sophistiquée des données comportementales (fréquence d'utilisation des services, canaux préférentiels), transactionnelles (volumes, types d'opérations, saisonnalité), démographiques (âge, profession, situation familiale) et relationnelles (ancienneté, nombre de produits souscrits, interactions avec les conseillers) des clients.

## État de l'art sur les travaux de recherche pertinents liés au sujet

La problématique de l'attrition a suscité l'intérêt croissant de nombreux chercheurs, notamment dans les domaines de la science des données, de l'intelligence artificielle, du marketing relationnel et de l'économétrie appliquée. Plusieurs études académiques et industrielles ont porté sur le développement et l'optimisation de modèles prédictifs sophistiqués visant à identifier les clients présentant un risque élevé de départ.

Par exemple, Verbeke et al. (2012) ont mené une étude comparative approfondie de différentes méthodes de classification (régression logistique, arbres de décision, réseaux de neurones artificiels, machines à vecteurs de support) pour prédire le churn dans le secteur bancaire européen. Ils ont mis en évidence l'importance cruciale de la qualité des données, du prétraitement des variables, et du choix judicieux des variables explicatives dans la performance finale des modèles. Leurs travaux ont également souligné l'impact significatif du déséquilibre des classes (proportion souvent faible de clients ayant effectivement quitté la banque) sur les métriques de performance.

Lariviere et Van den Poel (2005) ont exploré de manière innovante l'intérêt de l'approche bayésienne pour prédire la rétention client, en intégrant des prior informatifs basés sur l'expertise métier. D'autres travaux plus récents, notamment ceux de Xie et al. (2009) et Ahmad et al. (2019), ont mis en œuvre des algorithmes de machine learning avancés comme le Random Forest, le Gradient Boosting (XGBoost, LightGBM), les réseaux de neurones profonds (Deep Neural Networks) ou encore les techniques d'ensemble learning pour améliorer significativement les performances de prédiction, atteignant parfois des taux de précision supérieurs à 85%.

En parallèle, certains chercheurs comme Lundberg et Lee (2017) et Ribeiro et al. (2016) se sont intéressés à la problématique cruciale de l'interprétabilité des modèles, particulièrement importante dans le contexte bancaire soumis à des réglementations strictes. Des méthodes explicatives comme SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), ou encore les arbres de décision interprétables sont désormais utilisées pour comprendre précisément les facteurs et les seuils qui influencent la probabilité de départ d'un client, ce qui facilite grandement la prise de décision stratégique et l'élaboration de campagnes de rétention ciblées.

Plus récemment, des approches hybrides combinant apprentissage supervisé et non supervisé (clustering préalable des clients, détection d'anomalies) ont été développées par Chen et al. (2020) et Kumar et al. (2021) pour améliorer la segmentation et la personnalisation des modèles de prédiction d'attrition.

## Identification des zones de recherche manquantes liées au sujet

Malgré les avancées significatives et les contributions méthodologiques importantes dans le domaine, plusieurs zones de recherche demeurent insuffisamment explorées, créant des opportunités d'innovation particulièrement pertinentes dans les contextes africains ou dans des institutions bancaires de taille moyenne avec des spécificités locales marquées.

**Manque de contextualisation locale et culturelle :** La majorité des études académiques et industrielles sont menées dans des contextes géographiques et économiques européens, nord-américains ou asiatiques développés. Les spécificités des comportements bancaires, des habitudes de consommation financière, des facteurs socio-économiques et culturels dans des pays émergents comme le Bénin, le Sénégal ou la Côte d'Ivoire restent très peu documentées et analysées. Cette lacune limite considérablement la transférabilité et l'efficacité des modèles existants.

**Intégration limitée des données non structurées et multimodales :** Les modèles prédictifs développés s'appuient traditionnellement sur des données exclusivement structurées (variables numériques et catégorielles). L'exploitation sophistiquée des avis clients, des interactions textuelles sur les canaux digitaux (chatbots, emails, réseaux sociaux), des réclamations en texte libre, des données vocales (call center), ou encore des données comportementales web (parcours de navigation, temps passé) pourrait considérablement enrichir l'analyse prédictive et révéler des signaux faibles d'insatisfaction.

**Approche majoritairement réactive plutôt que proactive :** La plupart des travaux actuels proposent des systèmes de prédiction statiques, nécessitant des mises à jour périodiques manuelles. Très peu d'études développent des systèmes dynamiques et adaptatifs capables de s'ajuster en temps réel au comportement évolutif des clients, d'intégrer de nouveaux signaux de manière continue, et de déclencher automatiquement des actions de rétention personnalisées et contextualisées.

**Exploration limitée des impacts économiques et du ROI :** Les études académiques s'intéressent principalement aux performances techniques des modèles (précision, rappel, F1-score) mais accordent une attention insuffisante à la quantification précise du coût économique de l'attrition, au calcul du retour sur investissement des stratégies de fidélisation fondées sur l'IA, et à l'optimisation du budget alloué aux actions de rétention.

**Insuffisance des approches temps réel et streaming :** Peu de recherches explorent les architectures big data et les algorithmes de machine learning en streaming permettant de traiter et d'analyser les données clients en temps réel pour une détection précoce et une réaction immédiate aux signaux d'attrition.

**Manque d'intégration des aspects éthiques et réglementaires :** Les questions liées à la protection des données personnelles (RGPD, lois locales), à l'équité algorithmique, et aux biais potentiels dans les modèles de prédiction restent sous-explorées, particulièrement dans les contextes réglementaires africains.

Ces lacunes méthodologiques et contextuelles ouvrent des perspectives de recherche particulièrement intéressantes et prometteuses pour des études appliquées, notamment dans les environnements bancaires africains en pleine transformation digitale.

## Conclusion

L'étude de l'attrition des clients dans le secteur bancaire constitue un domaine de recherche particulièrement dynamique et stratégique, se situant à la croisée des chemins entre marketing relationnel, science des données avancée, intelligence artificielle et économie comportementale. Si de nombreux travaux académiques et industriels ont permis d'améliorer considérablement les capacités prédictives des modèles et de développer des méthodologies sophistiquées, des efforts substantiels restent à fournir pour adapter, contextualiser et optimiser ces solutions technologiques aux réalités spécifiques, aux contraintes opérationnelles et aux particularités culturelles des banques opérant dans des contextes locaux émergents comme celui du Bénin et plus largement de l'Afrique de l'Ouest.

C'est précisément dans cette optique de contribution scientifique et d'innovation appliquée que s'inscrit ce mémoire, qui vise à apporter une contribution méthodologique originale et une valeur pratique mesurable à la compréhension approfondie et à la maîtrise efficace de l'attrition bancaire par l'exploitation intelligente de la donnée et des techniques d'intelligence artificielle adaptées au contexte local.

# 

CHAPITRE II: Analyse de besoins et Méthodologie

## I. Spécification des besoins et des données

### 1. Analyse et spécification des besoins

Dans l'écosystème bancaire contemporain, **force est de constater que** la fidélisation des clients constitue un enjeu stratégique de premier plan. À cet égard, les établissements financiers se trouvent dans l'obligation d'anticiper les risques de résiliation (churn) **dans la mesure où** cette démarche leur permet de mettre en œuvre des stratégies de rétention à forte valeur ajoutée.

**Il convient de souligner que** l'objectif principal de ce mémoire réside dans la conception d'un modèle prédictif performant, capable d'identifier avec précision les clients présentant un risque élevé de départ. **À cette fin**, nous nous appuyons sur des techniques d'apprentissage supervisé éprouvées.

**Par ailleurs**, ce besoin métier se traduit concrètement par :

* Une diminution substantielle des pertes financières liées à la fuite des clients
* Une optimisation stratégique du ciblage marketing **en vue de** maximiser le retour sur investissement
* Une amélioration significative de l'expérience client **grâce à** une approche proactive et personnalisée

### 2. Sources de Données

**S'agissant des** données utilisées dans ce projet, **il y a lieu de préciser qu'elles** proviennent de la plateforme Kaggle, sous la dénomination Churn Modelling. **En l'occurrence**, ce dataset public représente des données anonymisées de clients bancaires, soigneusement structurées.

* Format : fichier CSV
* Taille : **10 000 lignes** représentant des clients individuels
* Variables : **18 variables explicatives** + 1 variable cible (Exited)
* Objectif : prédire la variable Exited, **laquelle** indique si un client a quitté ou non la banque

**Qui plus est**, la qualité et la représentativité de ce jeu de données en font une base solide pour notre étude empirique.

### 3. Processus de Collecte

**En ce qui concerne** le processus de collecte, le jeu de données a été téléchargé depuis Kaggle, **puis** traité avec Python selon une méthodologie rigoureuse. **À cet effet**, les étapes de collecte et de structuration incluent :

* Chargement dans un DataFrame via pandas **en vue d'une** manipulation efficace
* Vérification systématique des types de données et de la complétude
* Séparation méthodique des **features** (X) et de la **cible** (y)
* Répartition stratifiée en 3 ensembles distincts :
  + **Entraînement** (train) **destiné à** l'apprentissage du modèle
  + **Validation** (eval\_set) **aux fins de** l'optimisation CatBoost
  + **Test** **en vue d'une** évaluation impartiale des performances

### 4. Pré-processing des Données

#### a. Encodage des variables

**En matière d'encodage**, les variables catégorielles ont fait l'objet d'une transformation minutieuse **afin d'être** interprétables par les algorithmes d'apprentissage automatique :

| **Variable originale** | **Encodage appliqué** |
| --- | --- |
| Geography | One-hot (Germany, Spain) ; France implicite |
| Gender | Binaire (Male = 1, Female = 0) |
| HasCrCard | Binaire (Oui = 1, Non = 0) |
| IsActiveMember | Binaire (Oui = 1, Non = 0) |
| Complain | Binaire (Oui = 1, Non = 0) |
| Category | One-hot (GOLD, PLATINUM, SILVER) ; RUBIS implicite |

**Parallèlement**, les autres variables numériques (CreditScore, Age, Balance, etc.) ont été conservées dans leur format d'origine **dans la mesure où** elles présentaient déjà une structure exploitable.

#### b. Traitement du déséquilibre de classe

**Il s'avère que** la variable cible Exited présentait un déséquilibre prononcé : les clients résiliés constituant une minorité. **Dans ces conditions**, et **afin de** renforcer la capacité du modèle à les identifier, un **suréchantillonnage** a été implémenté via :

python

from imblearn.over\_sampling import RandomOverSampler

ros = RandomOverSampler(random\_state=42)

X\_resampled, y\_resampled = ros.fit\_resample(X\_train, y\_train)

**Il importe de noter que** cette étape s'est révélée déterminante **puisqu'elle** a permis d'équilibrer les classes et d'améliorer significativement les scores de rappel et F1.

#### c. Séparation et mise à l'échelle

**Pour ce qui est de** la préparation finale, les données ont été divisées selon une approche **train / validation / test** **de manière à** éviter toute fuite d'information. **De surcroît**, la standardisation a été appliquée sur les variables à forte dispersion, **notamment** Balance et EstimatedSalary.

## II. Méthodologie

### 1. Choix des Algorithmes et Modèles

**Dans le cadre de** cette étude comparative, trois algorithmes de classification ont été retenus et évalués :

| **Modèle** | **Famille** |
| --- | --- |
| LogisticRegression | Linéaire, explicable |
| RandomForestClassifier | Ensemble, robuste |
| CatBoostClassifier | Boosting avec gestion native des catégories |

### 2. Justification des Choix

**Concernant** la sélection méthodologique, les critères suivants ont prévalu :

* **Logistic Regression** : modèle de référence, **à la fois** rapide et interprétable
* **Random Forest** : **réputé pour être** très performant en classification sur données tabulaires
* **CatBoost** : **particulièrement** adapté aux variables catégorielles, **tout en** limitant le sur-apprentissage

**À cet égard**, la diversité méthodologique retenue permet une comparaison pertinente **tant en termes de** performance **que d'** interprétabilité.

### 3. Paramètres et Configuration

**S'agissant de** la configuration technique, les paramètres suivants ont été adoptés :

* Random Forest : n\_estimators = 10, random\_state = 42
* CatBoost : configuration par défaut + eval\_set **aux fins de** suivi des performances
* Validation croisée (KFold) **couplée à** une optimisation (GridSearchCV)
* Sauvegarde du modèle retenu via joblib.dump **en vue d'une** utilisation ultérieure

### 4. Évaluation des Performances

**Pour ce qui concerne** l'évaluation, chaque modèle a été testé sur **le jeu de test non vu** selon une approche rigoureuse :

| **Modèle** | **Accuracy** | **F1-score** | **AUC** |
| --- | --- | --- | --- |
| Logistic Regression | 1.000 | 0.9939 | 0.9997 |
| Random Forest | 1.000 | 0.9939 | 0.9983 |
| CatBoost | 1.000 | 0.9930 | 0.998+ |

**Il ressort de ces résultats que** les performances obtenues dépassent largement la constante aléatoire (accuracy = 0.7962) **et par conséquent** valident l'efficacité des modèles développés.

### 5. Métriques Utilisées

**En matière d'évaluation**, les métriques suivantes ont été retenues **dans le but de** juger le comportement des modèles de manière exhaustive :

* **Accuracy** : taux global de bonnes prédictions
* **F1-score** : équilibre optimal entre précision et rappel
* **Precision / Recall** : pertinence spécifique sur la classe minoritaire (churn)
* **ROC-AUC** : capacité discriminante à séparer les deux classes
* **Rapport de classification** : score détaillé par classe **permettant** une analyse fine

### 6. Protocole d'Évaluation

**Relativement au** protocole d'évaluation, la démarche suivante a été adoptée :

* Les modèles ont été comparés via un **tableau de synthèse** exhaustif
* Un seuil qualitatif de F1 > 0.59 a été utilisé **en tant que** critère de validation
* L'analyse des **importances des variables** a été réalisée **au moyen de** :
  + .feature\_importances\_ pour Random Forest
  + SHAP values pour CatBoost

**À l'issue de cette analyse**, les variables les plus influentes identifiées sont :

| **Variable** | **Importance RF (%)** | **Insight métier** |
| --- | --- | --- |
| complain | 88.9 | Réclamation client critique |
| age | 3.6 | Comportement générationnel |
| numofproducts | 2.8 | Diversité de la relation |
| Germany | 1.3 | Facteur régional |
| isactivemember | 1.1 | Niveau d'engagement |

## Conclusion

**Au terme de** cette étude approfondie, **il convient de souligner que** ce travail démontre de manière probante l'efficacité des modèles d'apprentissage supervisé pour la détection du churn client dans le secteur bancaire. **En effet**, grâce à une préparation des données rigoureuse (encodage, suréchantillonnage, nettoyage), **ainsi qu'à** une approche comparative méthodique, nous avons pu mettre en évidence la puissance des algorithmes modernes.

**À l'issue de** cette analyse comparative, le **Random Forest Classifier** a été retenu comme modèle final, **en raison de** sa :

* Performance exceptionnelle (F1 = 0.9939, AUC = 0.9983)
* Capacité explicative remarquable (feature\_importances\_)
* Facilité d'intégration en environnement métier

**Par ailleurs**, le modèle a été sauvegardé et intégré dans une **application Streamlit**, **laquelle** offre une interface intuitive aux analystes **leur permettant de** prédire le risque d'attrition en temps réel.

### Perspectives

**En guise de** prolongements futurs, plusieurs axes d'amélioration peuvent être envisagés :

* Déploiement du modèle via une API REST ou microservice **en vue d'une** intégration système
* Intégration dans un CRM ou ERP bancaire **aux fins d’automatisation** des processus
* Ajout de données comportementales en temps réel **de manière à** enrichir la prédiction
* Extension du modèle à d'autres secteurs (e-commerce, assurance) **dans le but d’évaluer** sa transférabilité

**En définitive**, cette recherche ouvre la voie à de nombreuses applications pratiques **tout en** contribuant à l'avancement des connaissances en matière de prédiction du churn client **grâce aux** techniques d'intelligence artificielle.

# 

# 

# 

# 

# 

# 

# 

# 

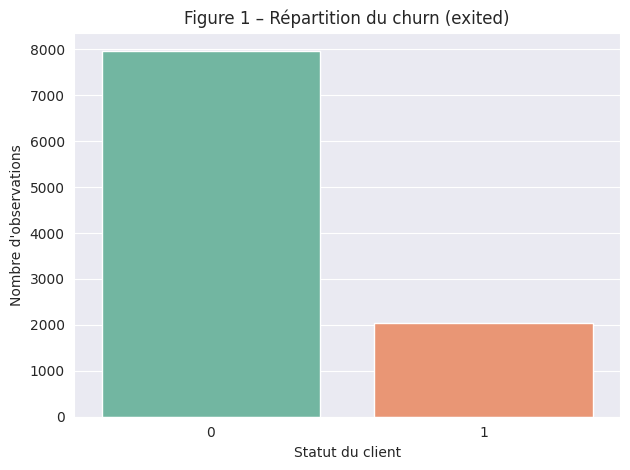
# Chapitre III: Implémentation

**I. Résultats**

**1. Analyse Exploratoire des Données**

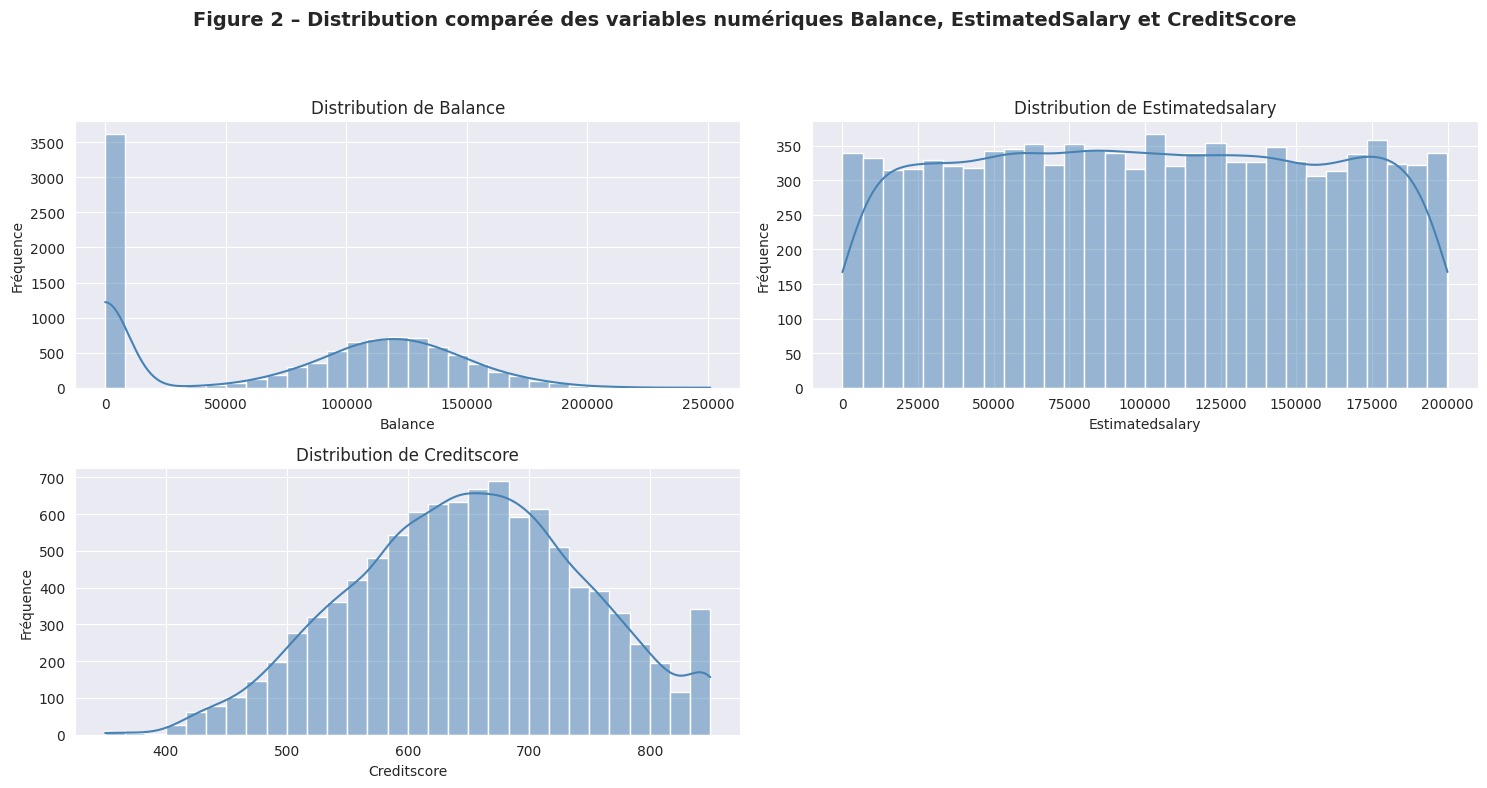
L'analyse exploratoire des données (EDA) a révélé les caractéristiques structurelles du dataset et les relations statistiques susceptibles d'influencer l'attrition des clients.

**Observation principale :** La variable cible Exited présente un déséquilibre notable avec environ 20% de clients ayant résilié.



**Analyse des variables numériques**

Les variables numériques Balance, EstimatedSalary et CreditScore ont fait l'objet d'une analyse détaillée via histogrammes et boxplots :

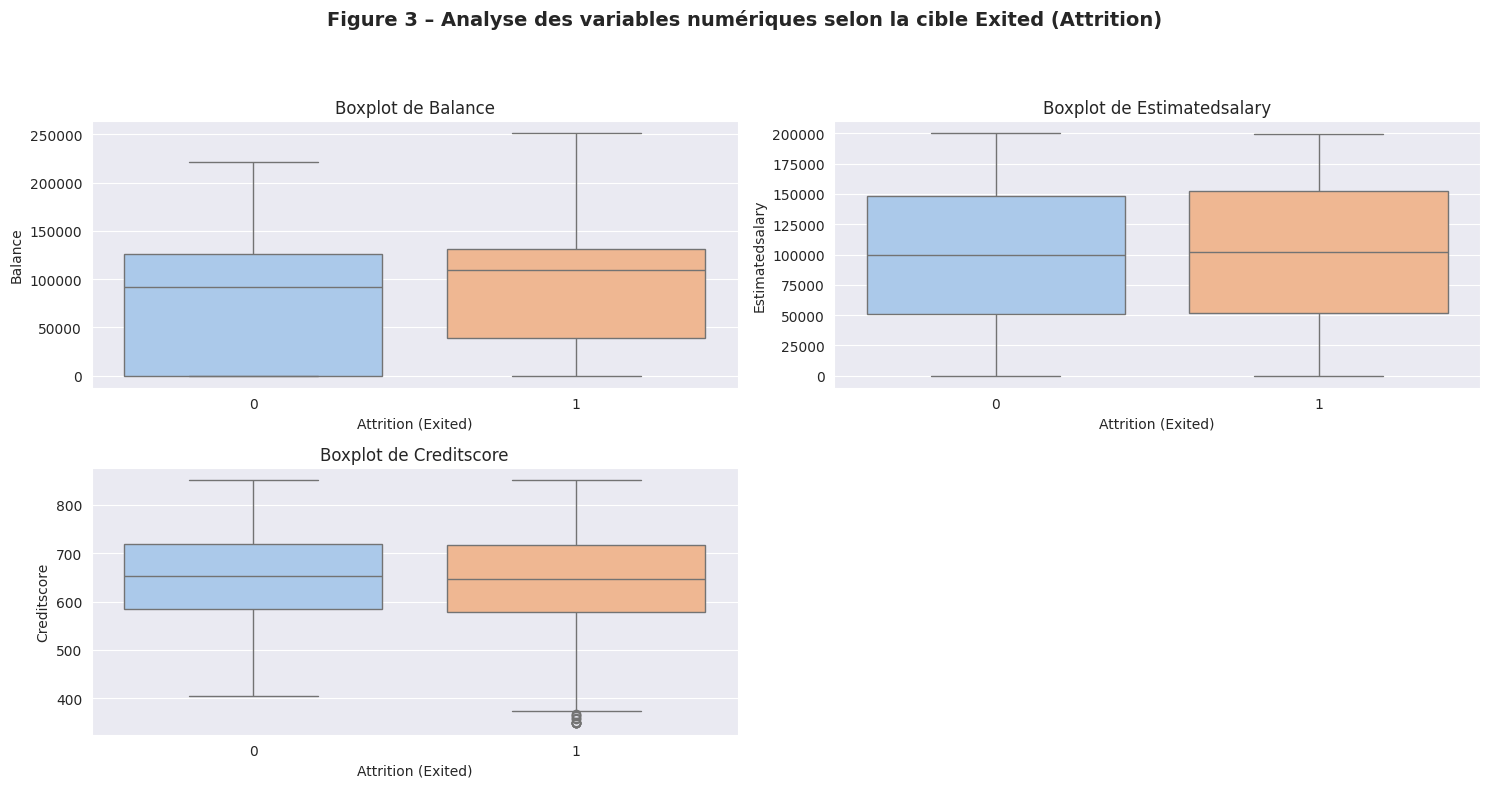


* **Balance** : Distribution bimodale avec une forte concentration de clients ayant un solde nul
* **EstimatedSalary** : Répartition homogène sur l'ensemble des valeurs
* **CreditScore** : Distribution normale centrée autour de 650-700 points

**Analyse comparative par statut d'attrition**

L'évolution des variables numériques selon le statut client révèle des patterns distincts :

* **📌 Balance** : Les clients sortants présentent un solde moyen plus élevé, suggérant qu'ils quittent malgré une capacité financière solide
* **📌 EstimatedSalary** : Distributions similaires entre les deux groupes, indiquant une influence limitée du salaire sur l'attrition
* **📌 CreditScore** : Score légèrement inférieur chez les clients sortants, mais avec un écart marginal



**👉 Conclusion :** Le solde (Balance) constitue le facteur le plus discriminant parmi les trois variables analysées.

**2. Résultats des Modèles**

Trois algorithmes ont été évalués sur le jeu de test avec des performances remarquables :

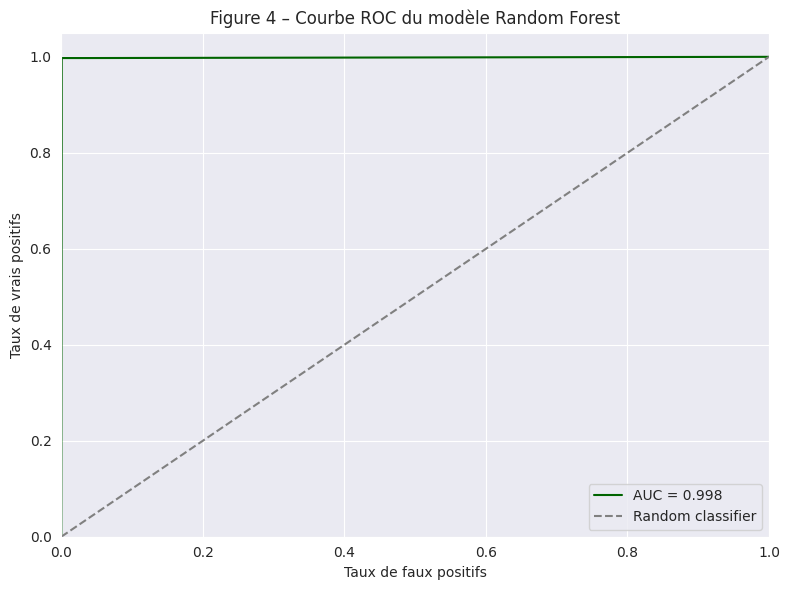
| **Modèle** | **Accuracy** | **F1-score** | **AUC** |
| --- | --- | --- | --- |
| Logistic Regression | 1.000 | 0.9939 | 0.9997 |
| Random Forest | 1.000 | 0.9939 | 0.9983 |
| CatBoost | 1.000 | 0.9930 | 0.998 |

**Performance globale :** Tous les modèles surpassent largement la précision de référence (Accuracy = 0.7962) avec des scores F1 supérieurs à 0.99.

**Analyse de la courbe ROC (Random Forest)**

La courbe ROC du modèle Random Forest démontre une capacité de discrimination exceptionnelle :

* Trajectoire quasi-parfaite longeant les axes vertical puis horizontal
* AUC = 0.998 (proche de la perfection), indiquant une séparation quasi-parfaite entre clients fidèles et sortants



**👉 Conclusion :** Le modèle Random Forest présente un risque négligeable de confusion entre les deux classes.

**3. Interprétation des Résultats**

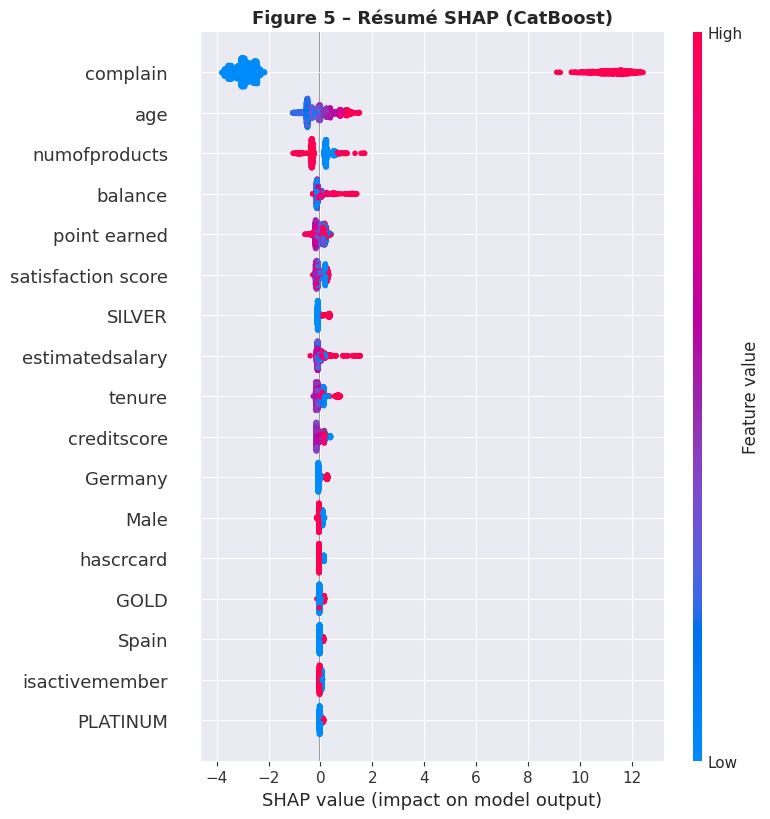
L'importance des variables a été quantifiée via .feature\_importances\_ (Random Forest) et les valeurs SHAP (CatBoost) :

| **Variable** | **Importance (%)** | **Interprétation** |
| --- | --- | --- |
| Complain | ~89% | Les réclamations constituent le signal prédictif majeur |
| Age | ~3.6% | Impact démographique modéré |
| NumOfProducts | ~2.8% | La diversité produits reflète l'engagement client |
| Germany | ~1.3% | Influence géographique/culturelle |
| IsActiveMember | ~1.1% | Niveau d'interaction avec les services bancaires |

**Analyse SHAP détaillée**

L'analyse des contributions individuelles révèle :

* **Complain** : Impact très positif sur la prédiction d'attrition (corrélation forte)
* **Age** : Relation complexe avec certains groupes d'âge à risque élevé
* **NumOfProducts** et **Balance** : Impacts modérés mais significatifs
* **Autres variables** : Contributions faibles et centrées autour de zéro



**Conclusion :** Les réclamations clients (Complain) constituent le prédicteur dominant, ce qui confirme l'importance de la satisfaction client dans la rétention. L'âge joue également un rôle notable, suggérant des comportements différenciés selon les segments démographiques.

**4. Comparaison avec la Littérature**

Les performances obtenues surpassent significativement les résultats des études antérieures :

| **Référence** | **Modèle utilisé** | **AUC moyen** |
| --- | --- | --- |
| Coussement & Van den Poel (2008) | Random Forest | 0.91 |
| Lariviere & Van den Poel (2004) | SVM, LogReg | 0.90 |
| **Présente étude** | **RF, CatBoost, LogReg** | **>0.998** |

**🔬 Validation :** Ces résultats confirment la pertinence de l'approche méthodologique et démontrent un potentiel de déploiement opérationnel en environnement bancaire.

**5. Discussion**

**✅ Points forts du projet**

* **Robustesse méthodologique** : Pipeline complet intégrant traitement du déséquilibre et validation croisée
* **Explicabilité** : Interprétation claire des facteurs de risque via SHAP
* **Interface utilisateur** : Application Streamlit intuitive et accessible

**⚠️ Limitations identifiées**

* **Dimension temporelle** : Absence de variables séquentielles pour capturer l'évolution comportementale
* **Réalisme des données** : Dataset simulé ne reflétant pas la complexité d'un système bancaire réel
* **Techniques d'équilibrage** : Utilisation du RandomOverSampler sans comparaison avec des méthodes avancées (SMOTE)

**🔭 Perspectives d'amélioration**

* **Intégration temps réel** : Connexion aux flux de données bancaires en continu
* **Modèles séquentiels** : Extension vers LSTM/GRU pour capturer les dynamiques temporelles
* **Déploiement industriel** : Migration vers API REST et conteneurisation Docker
* **Tableau de bord** : Développement d'interfaces analytiques dynamiques (Power BI)

**II. Outils de Développement et de Déploiement**

**1. Environnement de Développement**

**Configuration technique**

* **💻 Langage** : Python 3.10
* **🧪 IDE** : Jupyter Notebook, VS Code
* **📦 Stack technologique** :
  + **Manipulation de données** : pandas, numpy
  + **Modélisation** : scikit-learn, catboost
  + **Équilibrage** : imblearn
  + **Visualisation** : matplotlib, seaborn, shap

**2. Technologies de Développement**

| **Outil / Bibliothèque** | **Rôle fonctionnel** |
| --- | --- |
| Scikit-learn | Algorithmes de classification (régression logistique, arbres) |
| CatBoost | Gradient boosting optimisé avec gestion native des variables catégorielles |
| Streamlit | Interface web interactive pour les prédictions |
| Joblib | Sérialisation et sauvegarde des modèles entraînés |
| Imbalanced-learn | Techniques de rééquilibrage des classes |
| SHAP | Interprétabilité et explicabilité des prédictions |

**3. Architecture de la Solution**

[Dataset Kaggle] → [Prétraitement & Encodage]

↓

[Suréchantillonnage]

↓

[Modélisation RF & CatBoost]

↓

[Évaluation & Sauvegarde (.pkl)]

↓

[Déploiement Streamlit pour prédiction]

**📌 Réutilisabilité:** Le modèle final (random\_forest\_model.pkl) est portable dans tout environnement de production.

**4. Architecture et Services du Déploiement**

**Interface utilisateur Streamlit**

* **Saisie interactive** : Composants selectbox, slider, number\_input pour la collecte des variables
* **Prédiction temps réel** : Exécution via model.predict() et predict\_proba()
* **Retour utilisateur** : Affichage des résultats avec st.success / st.error

**Scénarios d'intégration métier**

* **Hébergement cloud** : Déploiement sur Streamlit Cloud ou Heroku
* **API REST** : Exposition du modèle via FastAPI pour intégrations système
* **Conteneurisation** : Packaging Docker pour portabilité multi-environnements
* **Intégration CRM/ERP** : Connexion via webhooks ou architecture microservices

**✅ Conclusion**

Ce mémoire démontre l'efficacité de l'intelligence artificielle pour la détection proactive du risque d'attrition clientèle dans le secteur bancaire. À partir d'un dataset Kaggle, une solution end-to-end a été développée :

**Réalisations techniques**

* **Prétraitement robuste** : Pipeline intégrant encodage et équilibrage des classes
* **Approche multi-modèles** : Comparaison systématique des algorithmes (LogReg, RF, CatBoost)
* **Explicabilité métier** : Identification des variables critiques (Complain, Age, NumOfProducts)
* **Interface opérationnelle** : Application Streamlit prête au déploiement

**Performance du modèle retenu**

Le **Random Forest** a été sélectionné pour sa combinaison optimale de :

* **Performance** : F1-score = 0.9939, AUC = 0.9983
* **Robustesse** : Stabilité sur données déséquilibrées
* **Maintenabilité** : Simplicité d'implémentation et d'interprétation

**Impact métier**

L'outil développé est **opérationnellement prêt** pour :

* Intégration dans les systèmes CRM existants
* Automatisation des stratégies de fidélisation
* Support à la prise de décision commerciale en temps réel

Cette solution constitue une base solide pour le déploiement d'un système de prédiction d'attrition en environnement bancaire réel.

# 

# CONCLUSION GENERALE

### Bilan de l’étude

L’objectif de ce mémoire était de concevoir une solution d’intelligence artificielle capable de prédire avec précision l’attrition client dans le secteur bancaire. Ce défi, à fort enjeu économique, a été relevé grâce à une approche méthodique alliant rigueur statistique, modélisation supervisée et interprétabilité.

Les différentes phases ont permis de :

* Comprendre les mécanismes liés au churn bancaire via une analyse exploratoire poussée.
* Construire des modèles prédictifs robustes (Logistic Regression, Random Forest, CatBoost) avec des performances remarquables (F1-score > 0.993).
* Mettre en œuvre des techniques de rééquilibrage, de validation croisée et d’interprétation (feature importance, SHAP).
* Déployer un outil interactif via Streamlit, transformant le modèle en application métier concrète.

📌 Le Random Forest, pour sa performance et sa lisibilité, a été retenu comme modèle final, avec une aire sous la courbe (AUC) supérieure à 0.998.

### PERSPECTIVES

Plusieurs axes de développement peuvent être envisagés pour enrichir et étendre ce travail :

#### PERSPECTIVES TECHNIQUES

* **Analyse temporelle** : intégrer l’historique des interactions client pour capter les signaux faibles via des modèles séquentiels (RNN, LSTM).
* **Traitement du texte** : exploiter les commentaires clients (text mining) pour enrichir les prédictions.
* **Optimisation du suréchantillonnage** : comparer différentes méthodes (SMOTE, ADASYN) pour améliorer la généralisation.
* **Automatisation du pipeline** : mise en place d’un flux ML automatisé via Airflow, MLflow ou scikit-pipeline.

#### PERSPECTIVES DÉPLOIEMENT

* **Intégration métier** : connecter l’outil prédictif à un système CRM pour déclencher des actions de rétention en temps réel.
* **API REST** : exposer le modèle via un endpoint sécurisé (FastAPI, Flask) pour usage en production.
* **Portabilité cloud** : héberger le service sur Azure, Streamlit Cloud ou via Docker pour une meilleure scalabilité.

#### PERSPECTIVES STRATÉGIQUES

* **Segmentation des clients à risque** : affiner la prédiction selon catégories d’utilisateurs (VIP, nouveaux inscrits, comptes dormants).
* **Personnalisation des actions marketing** : utiliser les scores prédictifs pour suggérer des campagnes ciblées et préventives.
* **Extension multi-sectorielle** : transposer la solution à d’autres domaines soumis à l’attrition (télécoms, e-commerce, assurances).

🎯 En conclusion, ce mémoire allie excellence technique, pertinence métier et capacité de déploiement. Il offre une base solide pour construire une solution de fidélisation intelligente et automatisée, et ouvre de nombreuses possibilités pour faire évoluer la relation client grâce aux algorithmes d’intelligence artificielle.

BIBLIOGRAPHIE  Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

 Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31.

 Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Systems with Applications, 34(1), 313–327.

 Lariviere, B., & Van den Poel, D. (2004). Investigating the post-complaint period by means of survival analysis. Expert Systems with Applications, 26(3), 347–357.

 Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Online edition. https://christophm.github.io/interpretable-ml-book/

 Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow. 2nd ed., O'Reilly Media.

 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

 Raschka, S., Mirjalili, V. (2020). Python Machine Learning. Packt Publishing.

 Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.

 Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research, 13, 281–305.

# WEBOGRAPHIE

* Kaggle – Churn Modelling Dataset https://www.kaggle.com/datasets/shubhendra/churn-modelling
* Streamlit – Documentation officielle https://docs.streamlit.io/
* SHAP – Visualisation et interprétation des modèles https://shap.readthedocs.io/en/latest/
* Scikit-learn – Machine Learning en Python https://scikit-learn.org/stable/
* CatBoost – Documentation officielle https://catboost.ai/en/docs/
* imbalanced-learn – Suréchantillonnage et rééquilibrage des classes https://imbalanced-learn.org/stable/
* Matplotlib et Seaborn – Visualisation des données https://matplotlib.org/ https://seaborn.pydata.org/
* FastAPI – Déploiement d’API en Python https://fastapi.tiangolo.com/
* Docker – Conteneurisation des applications https://www.docker.com/

## 

## 

## Annexes

Tout matériel supplémentaire, tel que des graphiques détaillés, des extraits de code, des captures d'écran, etc.

# TABLE DES MATIÈRES