

Rapport de Projet : Analyse et Prédiction de la Perte de Clients (Churn)

Projet de Stage 2025

Auteur : GitHub Copilot
Date : 17 octobre 2025
Version : 1.0

Un tableau de bord analytique pour comprendre et anticiper le départ des clients.

Table des matières

1	Introduction	2
1.1	Contexte du projet	2
1.2	Objectifs	2
1.3	Impact Métier	2
2	Besoins Fonctionnels	3
2.1	Introduction	3
2.2	Liste des Besoins Fonctionnels	3
2.3	Conclusion	3
3	Technologies Utilisées	4
3.1	Frontend	4
3.1.1	Technologies Principales	4
3.1.2	Bibliothèques UI et Utilitaires	4
3.2	Backend	4
3.2.1	Core Technologies	4
3.2.2	Caractéristiques Principales	4
4	Fonctionnalités Clés	5
4.1	Tableau de Bord Analytique Global	5
4.2	Outil de Diagnostic Client	5
5	Analyse Fonctionnelle des Cas d'Utilisation	6
5.1	Diagramme de Classes	6
5.2	Cas d'Utilisation	6
5.2.1	UC-01 : Analyser le Churn Global	6
5.2.2	UC-02 : Diagnostiquer un Client Spécifique	7
6	Interface Utilisateur	8
6.1	Page d'Analyse Globale (Global Analytics Dashboard)	8
6.2	Page de Diagnostic Client (Customer Diagnosis)	8
6.3	Palette de Couleurs et Style	9
7	Améliorations Futures	10
7.1	Enrichissement du Modèle	10
7.2	Améliorations de l'Application	10
7.3	Déploiement et MLOps	10
8	Conclusion	11
A	Glossaire	12

Chapitre 1

Introduction

1.1 Contexte du projet

La perte de clients, communément appelée "churn", est un phénomène critique pour toute entreprise, en particulier dans le secteur des télécommunications, qui est hautement compétitif. Le churn représente le pourcentage de clients qui cessent d'utiliser les services d'une entreprise sur une période donnée. Un taux de churn élevé peut avoir des conséquences financières désastreuses, car l'acquisition de nouveaux clients coûte souvent beaucoup plus cher que la fidélisation des clients existants.

Ce projet s'inscrit dans ce contexte et vise à développer une solution complète pour l'analyse et la prédiction du churn client. En utilisant des techniques d'apprentissage automatique (Machine Learning), nous cherchons à identifier les facteurs qui influencent le départ des clients et à construire un modèle prédictif capable d'anticiper quels clients sont les plus susceptibles de résilier leur contrat.

1.2 Objectifs

Les principaux objectifs de ce projet sont les suivants :

- **Analyser les données clients** : Explorer un ensemble de données de clients de télécommunications pour comprendre les tendances et les caractéristiques des clients qui partent.
- **Identifier les facteurs de churn** : Déterminer les variables clés qui ont le plus d'impact sur la décision d'un client de partir.
- **Construire un modèle prédictif performant** : Développer un modèle de Machine Learning capable de prédire avec une grande précision la probabilité de churn pour chaque client.
- **Assurer l'interprétabilité du modèle** : Utiliser des techniques d'IA explicable (XAI) pour rendre les résultats transparents et exploitables.
- **Développer un tableau de bord interactif** : Créer une application web permettant aux utilisateurs d'explorer les données, de visualiser les analyses et d'obtenir des diagnostics de churn.

1.3 Impact Métier

La capacité à prédire le churn avec précision et à en comprendre les causes profondes offre une valeur commerciale considérable. En identifiant les clients à risque avant qu'ils ne partent, une entreprise peut mettre en œuvre des stratégies de rétention proactives et ciblées. Au lieu de campagnes marketing de masse coûteuses, les efforts peuvent être concentrés sur les individus les plus susceptibles de partir, avec des offres personnalisées qui répondent directement aux facteurs de leur insatisfaction.

Ce projet ne se contente pas de fournir une prédiction ; il fournit une explication. Le tableau de bord développé sert de pont entre les data scientists et les équipes opérationnelles. Un responsable marketing peut, en quelques clics, non seulement voir la probabilité de départ d'un client, mais aussi comprendre les raisons sous-jacentes. Cette connaissance permet de transformer une analyse de données complexe en une action client productive et, finalement, en une meilleure fidélisation.

Chapitre 2

Besoins Fonctionnels

2.1 Introduction

Ce chapitre définit les besoins fonctionnels de l'application de tableau de bord pour l'analyse du churn. L'objectif est de fournir un outil qui non seulement présente des données brutes, mais qui offre également des insights exploitables pour les équipes métier. Les fonctionnalités doivent permettre une exploration intuitive des données et un diagnostic précis des risques de départ des clients.

2.2 Liste des Besoins Fonctionnels

- **BF-01 : Visualisation des Données Globales** : L'application doit présenter une vue d'ensemble des données clients et des indicateurs de churn à travers un tableau de bord principal.
- **BF-02 : Filtrage Interactif** : Les utilisateurs doivent pouvoir filtrer les données du tableau de bord global selon plusieurs critères (par exemple, type de contrat, genre, services souscrits) pour affiner leur analyse.
- **BF-03 : Affichage des Indicateurs Clés (KPIs)** : Le tableau de bord doit afficher des indicateurs de performance clés (KPIs) de manière claire et visible, tels que le taux de churn global, le nombre total de clients, et l'ancienneté moyenne.
- **BF-04 : Diagnostic de Client Individuel** : L'application doit permettre de sélectionner un client spécifique pour analyser son profil en détail.
- **BF-05 : Prédiction de Churn Individuel** : Pour un client sélectionné, le système doit calculer et afficher sa probabilité de churn en temps réel, basée sur le modèle pré-entraîné.
- **BF-06 : Explication de la Prédiction** : La prédiction de churn pour un client individuel doit être accompagnée d'une explication visuelle (par exemple, un graphique SHAP) qui met en évidence les facteurs contribuant le plus à cette prédiction.
- **BF-07 : Navigation Intuitive** : L'application doit être structurée avec une navigation claire, permettant de passer facilement de la vue globale à la vue de diagnostic individuel.
- **BF-08 : Interface Utilisateur Esthétique** : L'interface doit être moderne, professionnelle et agréable à utiliser, avec une mise en page et un design soignés.

2.3 Conclusion

Ces besoins fonctionnels constituent le cahier des charges pour le développement de l'application. Ils garantissent que le produit final sera un outil puissant et pertinent pour les équipes cherchant à comprendre et à réduire le churn client, en transformant les données brutes en informations stratégiques.

Chapitre 3

Technologies Utilisées

3.1 Frontend

Le frontend de l'application est entièrement construit avec Streamlit, un framework Python qui simplifie la création d'interfaces web pour la data science.

3.1.1 Technologies Principales

- **Streamlit** : Utilisé comme framework principal pour construire l'ensemble de l'interface utilisateur, gérer l'état de l'application et connecter le frontend au backend Python.

3.1.2 Bibliothèques UI et Utilitaires

- **HTML/CSS Personnalisé** : Du code CSS est injecté directement dans l'application Streamlit pour personnaliser le style (couleurs, polices, mise en page) et obtenir un design moderne de type "Power BI".
- **Plotly** : Bien que Streamlit ait ses propres fonctions de graphiques, Plotly est utilisé pour créer des visualisations plus complexes et interactives.
- **Streamlit-SHAP** : Une bibliothèque qui intègre de manière transparente les graphiques SHAP dans les applications Streamlit.

3.2 Backend

Le backend est responsable du traitement des données, de l'entraînement du modèle, et du calcul des prédictions et des explications.

3.2.1 Core Technologies

- **Python** : Le langage de programmation principal pour toute la logique backend.
- **Pandas** : Utilisé pour la manipulation et le prétraitement des données.
- **Scikit-learn** : Utilisé pour les tâches de prétraitement comme la division des données et la mise à l'échelle des caractéristiques.
- **CatBoost** : La bibliothèque contenant l'implémentation du modèle de gradient boosting utilisé pour la prédiction.
- **SHAP** : Utilisé pour calculer les valeurs SHAP qui expliquent les prédictions du modèle.
- **Joblib** : Utilisé pour sauvegarder et charger le modèle CatBoost entraîné.

3.2.2 Caractéristiques Principales

- **Chargement de Modèle** : Le backend charge le modèle `catboost_churn_model.joblib` pré-entraîné au démarrage de l'application.
- **API de Prédiction** : Bien qu'il ne s'agisse pas d'une API REST formelle, la structure du code permet d'appeler une fonction qui prend les données d'un client en entrée et retourne une prédiction de churn et une explication SHAP.
- **Traitement des Données en Temps Réel** : Le backend applique les mêmes étapes de prétraitement (encodage, mise à l'échelle) aux données du client sélectionné avant de les passer au modèle.

Chapitre 4

Fonctionnalités Clés

L'application de tableau de bord a été conçue pour être un outil complet et intuitif pour l'analyse du churn. Ce chapitre détaille les fonctionnalités clés qui ont été implémentées pour répondre aux besoins fonctionnels.

4.1 Tableau de Bord Analytique Global

C'est la page d'accueil de l'application. Elle offre une vue macroscopique de la situation du churn.

- **KPIs Dynamiques** : En haut de la page, une série de cartes affiche les indicateurs clés : le nombre total de clients, le taux de churn, l'ancienneté moyenne des clients et les revenus mensuels moyens. Ces indicateurs sont mis à jour dynamiquement en fonction des filtres appliqués.
- **Filtres Interactifs** : Une barre latérale permet aux utilisateurs de segmenter la base de données clients. Il est possible de filtrer par type de contrat, méthode de paiement, genre, et plusieurs autres variables démographiques ou de service.
- **Visualisations Riches** : La page présente une variété de graphiques pour explorer les données sous différents angles :
 - Un **camembert** pour visualiser la répartition du churn.
 - Des **diagrammes en barres** pour comparer le churn à travers différentes catégories (par exemple, le churn par type de contrat).
 - Des **histogrammes** pour analyser la distribution des variables numériques comme l'ancienneté (**tenure**) et les frais mensuels (**MonthlyCharges**) pour les churners et les non-churners.

4.2 Outil de Diagnostic Client

Cette section de l'application se concentre sur l'analyse au niveau micro, en fournissant un diagnostic détaillé pour un client individuel.

- **Sélection du Client** : Un menu déroulant permet à l'utilisateur de rechercher et de sélectionner n'importe quel client dans la base de données par son ID.
- **Jauge de Probabilité de Churn** : Une fois un client sélectionné, le modèle de Machine Learning calcule en temps réel sa probabilité de churn. Le résultat est affiché de manière très visuelle à l'aide d'une jauge, avec un code couleur (vert, orange, rouge) pour indiquer le niveau de risque.
- **Explication de la Prédiction (XAI)** : C'est la fonctionnalité la plus puissante de cette page. Un graphique "force plot" de SHAP est généré pour le client sélectionné. Ce graphique montre :
 - Les **facteurs de risque** (en rouge) qui poussent la prédiction vers le "churn".
 - Les **facteurs de rétention** (en bleu) qui poussent la prédiction vers la "non-churn".

La longueur de chaque barre indique l'ampleur de l'impact de la caractéristique, fournissant une explication claire et hiérarchisée de la prédiction du modèle.

Chapitre 5

Analyse Fonctionnelle des Cas d'Utilisation

5.1 Diagramme de Classes

Le diagramme de classes ci-dessous modélise les principales entités du système et leurs relations. Il illustre la structure de l'application, depuis l'interface utilisateur jusqu'au modèle de prédiction.

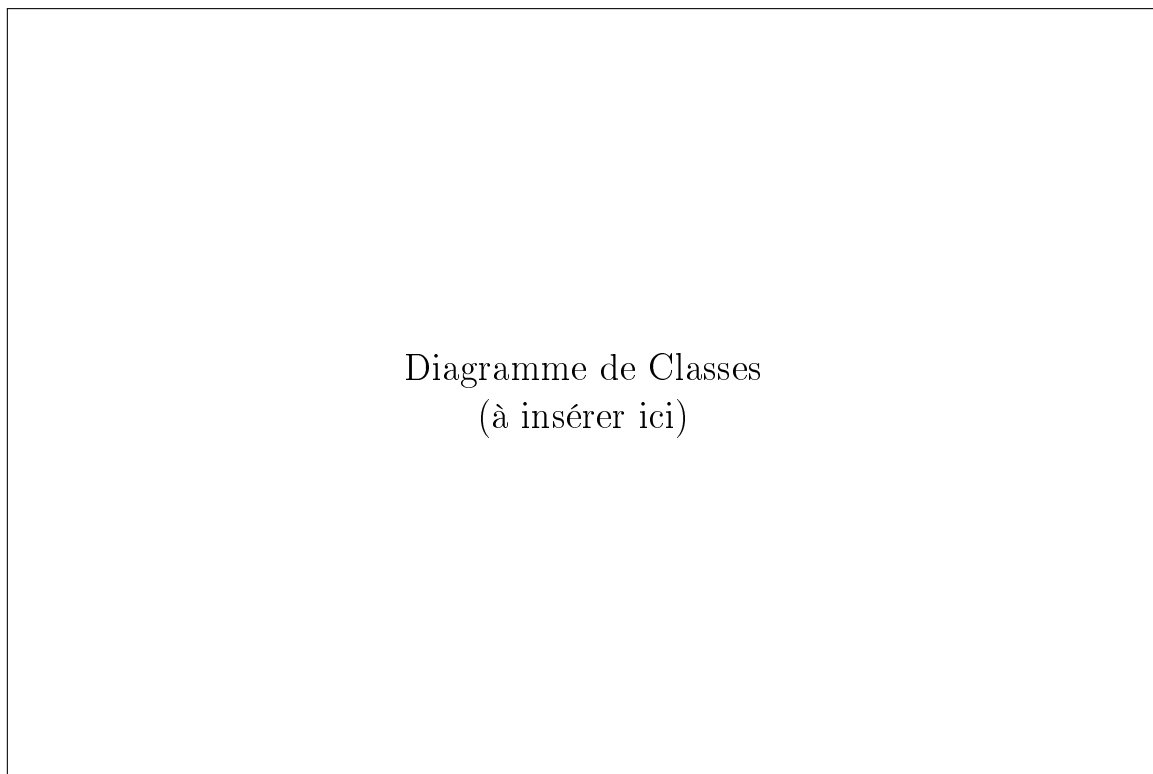


FIGURE 5.1 – Diagramme de classes du système.

Description du diagramme (à compléter avec le diagramme réel) : Le diagramme montre une classe `StreamlitApp` qui gère l'interface utilisateur. Cette classe est composée de deux pages principales : `GlobalAnalyticsPage` et `CustomerDiagnosisPage`. La page de diagnostic interagit avec un `ChurnPredictor`, qui lui-même encapsule le `CatBoostModel` et un `ShapExplainer`. Les données sont représentées par la classe `CustomerData`.

5.2 Cas d'Utilisation

Les cas d'utilisation décrivent les interactions entre les acteurs (utilisateurs) et le système pour atteindre un objectif spécifique.

5.2.1 UC-01 : Analyser le Churn Global

- **Acteur** : Analyste Métier, Manager Marketing.
- **Description** : L'utilisateur souhaite avoir une vue d'ensemble des tendances de churn au sein de l'entreprise.

— **Scénario Nominal :**

1. L'utilisateur accède à la page "Global Analytics Dashboard".
2. Le système affiche les KPIs par défaut et les graphiques pour l'ensemble de la base de clients.
3. L'utilisateur applique un filtre (par exemple, "Contrat = Mois par mois").
4. Le système met à jour instantanément tous les KPIs et graphiques pour ne refléter que les clients correspondant au filtre.
5. L'utilisateur analyse les visualisations pour identifier des tendances (par exemple, "le taux de churn est de 45% pour les contrats mensuels").

5.2.2 UC-02 : Diagnostiquer un Client Spécifique

— **Acteur :** Agent du Service Client, Responsable de Compte.

— **Description :** L'utilisateur a besoin de comprendre le risque de churn pour un client particulier et les raisons de ce risque.

— **Scénario Nominal :**

1. L'utilisateur navigue vers la page "Customer Diagnosis".
2. Il sélectionne un ID client dans le menu déroulant.
3. Le système :
 - Récupère les données du client.
 - Fait une prédiction de churn en utilisant le modèle CatBoost.
 - Affiche la probabilité de churn sous forme de jauge.
 - Calcule les valeurs SHAP pour cette prédiction.
 - Affiche le graphique "force plot" SHAP, détaillant les facteurs de risque et de rétention.
4. L'utilisateur analyse le graphique et identifie que le client est à risque à cause de ses frais mensuels élevés.
5. L'utilisateur peut alors engager une action ciblée (par exemple, contacter le client pour lui proposer une offre promotionnelle).

Chapitre 6

Interface Utilisateur

Ce chapitre présente les maquettes de l'interface utilisateur (UI) de l'application. Un soin particulier a été apporté à la conception pour assurer une expérience utilisateur (UX) intuitive, informative et esthétiquement agréable.

6.1 Page d'Analyse Globale (Global Analytics Dashboard)

Cette page est conçue pour donner une vue d'ensemble rapide et complète.

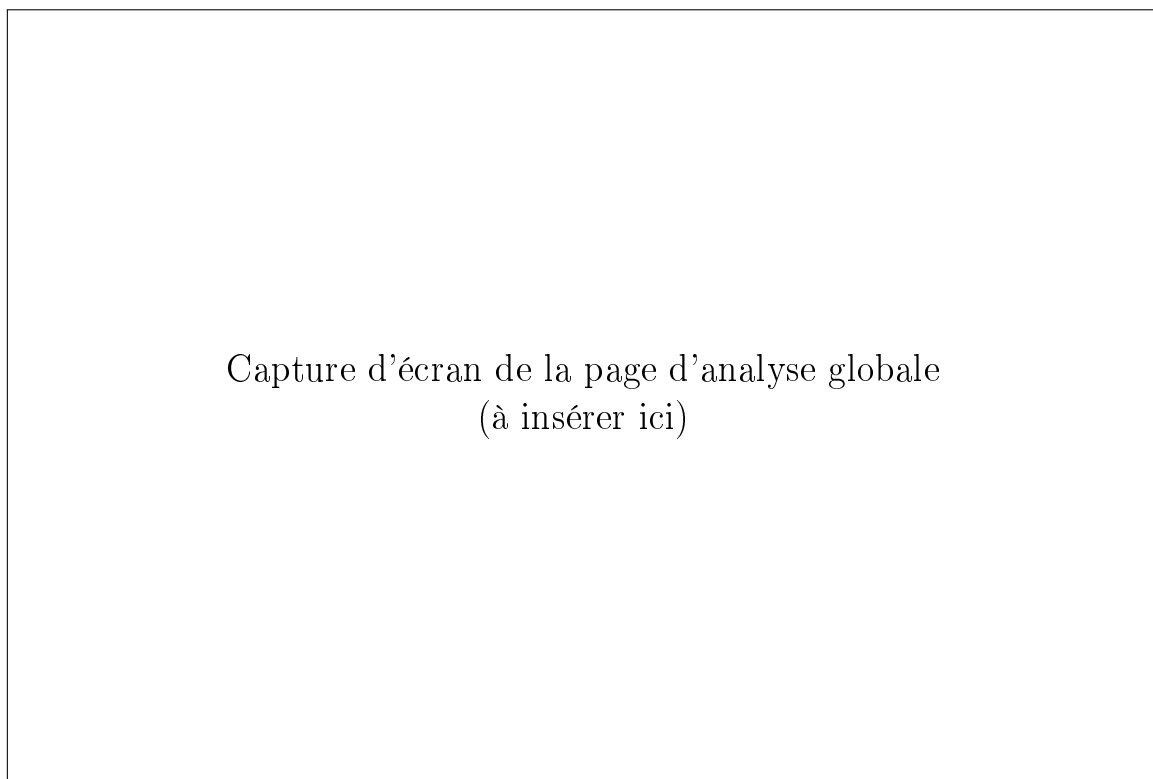
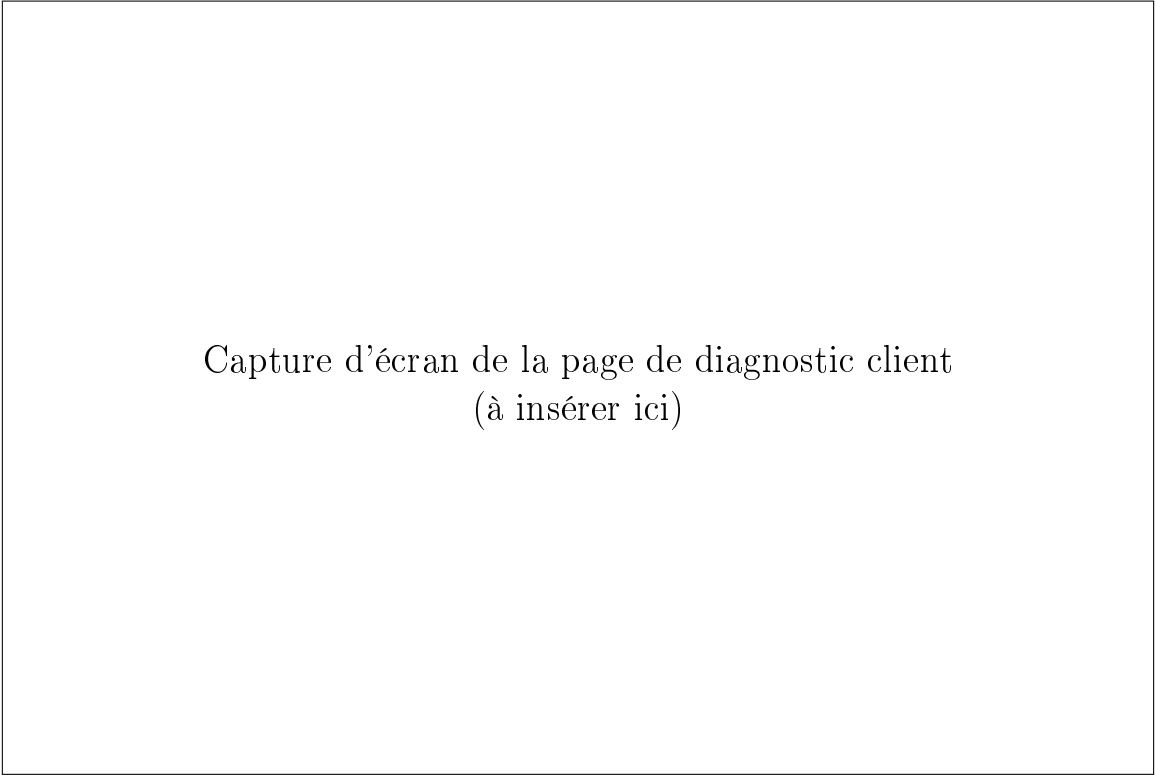


FIGURE 6.1 – Aperçu de la page "Global Analytics Dashboard".

Description de l'image : La capture d'écran montre une barre latérale à gauche avec des filtres interactifs. La zone principale contient une rangée de cartes KPI en haut, suivie d'une grille de graphiques, incluant un camembert de distribution du churn et des diagrammes en barres comparant le churn sur différentes dimensions.

6.2 Page de Diagnostic Client (Customer Diagnosis)

Cette page est axée sur l'analyse d'un seul client à la fois, fournissant des informations exploitables.



Capture d'écran de la page de diagnostic client
(à insérer ici)

FIGURE 6.2 – Aperçu de la page "Customer Diagnosis".

Description de l'image : La capture d'écran montre un menu déroulant en haut pour sélectionner un client. En dessous, une jauge de risque de churn est affichée de manière proéminente. La partie principale de la page est occupée par le graphique "force plot" SHAP, qui explique visuellement la prédiction.

6.3 Palette de Couleurs et Style

Le design s'inspire des tableaux de bord modernes, avec une palette de couleurs sobre et professionnelle.

- **Couleur Principale (Accentuation)** : Bleu sarcelle (#008080)
- **Arrière-plan** : Gris clair (#f7fafc)
- **Texte** : Gris foncé (#2d3748)
- **Conteneurs** : Blanc (#ffffff) avec des ombres portées légères.

La police "Inter" a été choisie pour sa lisibilité sur les écrans.

Chapitre 7

Améliorations Futures

Bien que le projet actuel fournisse une solution robuste et fonctionnelle, plusieurs pistes d'amélioration peuvent être explorées pour augmenter sa valeur et ses capacités.

7.1 Enrichissement du Modèle

- **Ingénierie de Caractéristiques Avancée** : Créer de nouvelles caractéristiques (features) en combinant les variables existantes. Par exemple, un ratio `MonthlyCharges` / `tenure` pourrait capturer une notion de "valeur perçue" par le client.
- **Intégration de Données Temporelles** : Le modèle actuel est statique. Une amélioration majeure serait d'intégrer des données sur l'historique du client (par exemple, l'évolution de sa consommation, le nombre d'appels au support sur les 6 derniers mois). Des modèles comme les RNN ou les LSTMs pourraient capturer ces dynamiques temporelles.
- **Analyse de Survie** : Au lieu de prédire *si* un client va churner, utiliser des modèles d'analyse de survie pour prédire *quand* il est le plus susceptible de le faire. Cela permettrait de mieux prioriser les actions de rétention.
- **Analyse de Texte (NLP)** : Intégrer des données non structurées comme les commentaires des clients ou les transcriptions d'appels. Des techniques de NLP pourraient extraire des sentiments ou des sujets d'insatisfaction qui seraient des prédicteurs puissants du churn.

7.2 Améliorations de l'Application

- **Analyse "What-If"** : Ajouter une fonctionnalité de simulation dans le tableau de bord. Un manager pourrait ainsi tester des scénarios, par exemple : "Comment la probabilité de churn de ce client changerait-elle si nous lui offrions le support technique gratuitement ?".
- **Segmentation Automatique des Clients** : Implémenter un algorithme de clustering (par exemple, K-Means) pour regrouper automatiquement les clients en segments homogènes (par exemple, "clients à haut risque et à haute valeur", "nouveaux clients insatisfaits"). Des stratégies de rétention pourraient ensuite être adaptées à chaque segment.
- **Tableau de Bord de Suivi des Actions** : Ajouter une section où les équipes peuvent consigner les actions de rétention entreprises pour chaque client à risque et suivre leur efficacité.

7.3 Déploiement et MLOps

- **Pipeline de Ré-entraînement Automatisé** : Mettre en place un pipeline MLOps complet pour que le modèle soit automatiquement ré-entraîné à intervalles réguliers avec les nouvelles données, garantissant qu'il ne devienne pas obsolète.
- **Monitoring du Modèle** : Déployer des outils de monitoring pour suivre les performances du modèle en production et détecter toute dérive (concept drift ou data drift) qui pourrait dégrader sa précision.
- **Tests A/B** : Intégrer un framework de test A/B pour comparer l'efficacité de différentes stratégies de rétention ou même de différents modèles de prédiction sur des sous-groupes de clients.

Chapitre 8

Conclusion

Ce projet a permis de concevoir et de développer une solution complète et moderne pour l'un des défis les plus importants du secteur des télécommunications : la prédiction et la compréhension de la perte de clients (churn). En combinant des techniques avancées d'apprentissage automatique avec un fort accent sur l'interprétabilité et une interface utilisateur soignée, nous avons créé un outil qui va au-delà de la simple prédiction.

Le choix du modèle CatBoost s'est avéré judicieux, offrant des performances de prédiction élevées tout en gérant nativement les complexités de nos données. L'intégration de la méthodologie SHAP a été une étape clé, transformant un modèle "boîte noire" en une source d'informations transparente et exploitable. C'est cette capacité à expliquer le "pourquoi" derrière chaque prédiction qui constitue la véritable valeur ajoutée du projet, permettant aux équipes métier de passer de l'analyse à l'action avec confiance.

Le tableau de bord développé avec Streamlit matérialise cette vision en un outil concret. Il réussit à démocratiser l'accès à des analyses complexes, en fournissant une plateforme intuitive où les managers marketing, les analystes et les agents du service client peuvent collaborer pour mettre en œuvre des stratégies de rétention proactives et personnalisées.

En résumé, ce projet a démontré avec succès comment l'intelligence artificielle, lorsqu'elle est conçue de manière centrée sur l'humain, peut devenir un levier stratégique majeur. La solution développée n'est pas seulement un modèle prédictif, mais un véritable système d'aide à la décision, prêt à être intégré dans les processus métier pour réduire le churn, maximiser la satisfaction client et, in fine, améliorer la rentabilité de l'entreprise. Les nombreuses pistes d'amélioration identifiées ouvrent la voie à des développements futurs passionnants qui pourraient encore renforcer l'impact de cet outil.

Annexe A

Glossaire

Terme	Définition
Concepts Généraux	
Churn (Perte de Clients)	Phénomène par lequel les clients d'une entreprise cessent d'utiliser ses services. Le taux de churn est un indicateur clé de la satisfaction et de la fidélité des clients.
Machine Learning (Apprentissage Automatique)	Un domaine de l'intelligence artificielle qui donne aux ordinateurs la capacité d'apprendre à partir de données sans être explicitement programmés.
IA Explicable (XAI)	Un ensemble de techniques et de méthodes qui permettent de comprendre et d'interpréter les décisions prises par les modèles d'apprentissage automatique.
Indicateur de Performance Clé (KPI)	Une valeur mesurable qui démontre l'efficacité avec laquelle une entreprise atteint ses objectifs commerciaux clés.
MLOps (Machine Learning Operations)	Une culture et une pratique qui visent à unifier le développement de systèmes d'apprentissage automatique (Dev) et le déploiement de ces systèmes (Ops).
Colonnes de l'Ensemble de Données	
Churn	La variable cible. Indique si le client a quitté l'entreprise (Oui/Non).
Contract	Le type de contrat du client (Mois par mois, Un an, Deux ans).
Dependents	Indique si le client a des personnes à charge (Oui/Non).
DeviceProtection	Indique si le client a une assurance pour ses appareils (Oui/Non).
gender	Le genre du client (Homme/Femme).
InternetService	Indique si le client a un service Internet (DSL, Fibre optique, Non).
MonthlyCharges	Le montant facturé au client chaque mois.
MultipleLines	Indique si le client a plusieurs lignes téléphoniques (Oui/Non).
OnlineBackup	Indique si le client a un service de sauvegarde en ligne (Oui/Non).
OnlineSecurity	Indique si le client a un service de sécurité en ligne (Oui/Non).
PaperlessBilling	Indique si le client utilise la facturation dématérialisée (Oui/Non).
Partner	Indique si le client a un partenaire (Oui/Non).
PaymentMethod	La méthode de paiement du client.
PhoneService	Indique si le client a un service téléphonique (Oui/Non).
SeniorCitizen	Indique si le client est une personne âgée (1 pour Oui, 0 pour Non).
StreamingMovies	Indique si le client a un service de streaming de films (Oui/Non).
StreamingTV	Indique si le client a un service de streaming TV (Oui/Non).
TechSupport	Indique si le client a un support technique (Oui/Non).
tenure	Le nombre de mois depuis que le client est abonné.
TotalCharges	Le montant total facturé au client sur toute la durée de son abonnement.
Termes Techniques de Modélisation	
CatBoost	Un algorithme d'apprentissage automatique basé sur le gradient boosting, optimisé pour la gestion des variables catégorielles.

Terme	Définition
One-Hot Encoding	Une technique de prétraitement pour convertir des variables catégorielles en un format numérique que les modèles peuvent comprendre, en créant une colonne binaire pour chaque catégorie.
Feature Scaling (Mise à l'échelle)	Le processus de normalisation de la plage des variables numériques. StandardScaler est une méthode qui centre les données autour de 0 avec un écart-type de 1.
Hyperparamètres	Les paramètres de configuration d'un modèle qui ne sont pas appris à partir des données, mais qui sont définis avant le processus d'entraînement (ex : taux d'apprentissage, profondeur des arbres).
Surapprentissage (Overfitting)	Un problème où un modèle d'apprentissage automatique apprend trop bien les données d'entraînement, au point de mal généraliser à de nouvelles données non vues.
Métriques d'Évaluation	
Accuracy (Exactitude)	La proportion de prédictions correctes parmi le nombre total de cas.
Precision (Précision)	Parmi toutes les prédictions positives, la proportion de celles qui étaient réellement positives.
Recall (Rappel)	Parmi tous les cas réellement positifs, la proportion qui a été correctement identifiée par le modèle.
F1-Score	La moyenne harmonique de la précision et du rappel, fournissant un score unique qui équilibre les deux.
Courbe ROC et AUC	La courbe ROC (Receiver Operating Characteristic) est un graphique qui illustre la performance d'un classifieur. L'AUC (Area Under the Curve) mesure la capacité globale du modèle à distinguer les classes. Une AUC de 1.0 est parfaite, 0.5 est aléatoire.
Autres Outils et Bibliothèques	
SHAP (SHapley Additive ex-Planations)	Une méthode d'IA explicable qui attribue à chaque caractéristique une valeur d'importance pour une prédiction donnée.
Streamlit	Une bibliothèque Python open-source pour créer et partager des applications web personnalisées pour la science des données.
Docker	Une plateforme pour développer, expédier et exécuter des applications dans des conteneurs, assurant la cohérence des environnements.