

Parcours de Formation Experts Intelligence Artificielle

CGI – DAUPHINE

PROJET FINAL

PREVENTION DE L'ATTRITION (CHURN) DANS LE SECTEUR BANCAIRE

ABOUELFADL Khaoula • **FERNANDES BORGES** Manuel • **OUARDANI** Ismail

Table des matières

1	Définition du problème et objectifs	4
1.1	Problème	4
1.1.1	Comprendre le churn pour mieux le combattre	4
1.1.2	Identifier les formes de churn pour affiner les stratégies	4
1.1.3	Mesurer le churn : évaluer l'ampleur des pertes	4
1.1.4	Maîtriser le churn : les armes du stratège	5
1.1.5	Le churn : un indicateur clé de la stratégie d'entreprise	5
1.1.6	Prévenir le churn : l'écoute, la clé de la fidélisation	5
1.2	Objectifs	5
2	Gestion des données	6
2.1	Collecte et pré-traitement	6
2.1.1	Importation et Examen des Données	6
2.1.2	Analyse de la Distribution des Scores de Crédit	6
2.1.3	Identification des Facteurs de Churn	7
2.1.4	Analyse des corrélations	7
2.1.5	Préparation pour la modélisation	8
2.2	Qualité des données	8
2.2.1	Préparation des données	8
2.2.2	Normalisation des données	8
3	Méthodologie	9
3.1	Choix des algorithmes	9
3.1.1	Mise en Place et Test des Algorithmes	9
3.1.2	Sélection des caractéristiques	9
3.1.3	Choix du modèle : test et évaluation	9
3.1.4	Modèle retenu : arbre décisionnel	9
3.1.5	Conclusion sur le modèle	10
3.2	Approche théorique (justification théorique des choix méthodologiques)	10
3.2.1	Identification du problème d'Overfitting	10
3.2.2	Stratégies de gestion de l'Overfitting	11
	a. Élagage de l'arbre de décision (Pruning)	11
	b. Validation croisée (Cross-Validation)	11
3.2.3	Conclusion	11

4	Performance du modèle	12
4.1	Amélioration du modèle post-Overfitting	12
4.2	Résultats améliorés	12
4.3	Identification des variables influentes	12
4.4	Conclusion	13
5	Perspectives	13
5.1	Perspectives et recommandations business	13
6	MLOps	14
6.1	Déploiement	14
6.2	Surveillance et maintenance	14
6.2.1	Mise en place d'une métrique de performance de référence	14
6.2.2	Surveillance continue et tests de non-régression	15
6.2.3	Gestion du concept drift	15
6.2.4	Versionnement des modèles et rollback	16
6.2.5	Enrichissement et réentraînement des modèles	16
6.2.6	Outils et Infrastructure	16

1 Définition du problème et objectifs

1.1 Problème

Le churn, contraction des termes "change" et "turn", représente une menace insidieuse pour toute entreprise. Il incarne le désaveu des clients, une fuite qui, si elle n'est pas maîtrisée, peut ébranler même les structures les plus robustes.

1.1.1 Comprendre le churn pour mieux le combattre

Le churn, ou taux d'attrition, est un fléau dont il faut saisir les nuances avant de pouvoir l'endiguer. Il quantifie, en pourcentage, la perte de clients, d'utilisateurs ou d'abonnés. Longtemps confiné aux secteurs de la finance et des télécommunications, il touche aujourd'hui tous les domaines.

Il est essentiel de distinguer deux formes de churn : volontaire et involontaire. Le churn volontaire survient lorsqu'un client, déçu, abandonne une entreprise après une expérience négative. Cette rupture est souvent irrévocable. Le churn involontaire, quant à lui, est lié à des facteurs externes, tels qu'un déménagement qui impose l'annulation d'un service.

1.1.2 Identifier les formes de churn pour affiner les stratégies

Pour élaborer des stratégies de rétention efficaces, il est crucial de reconnaître les trois manifestations du churn :

1. **L'abandon** : Un client cesse d'utiliser un produit ou service, préférant une alternative (par exemple, quitter un abonnement aux transports en commun pour adopter le vélo). Cela témoigne d'une désaffection complète.

2. **Le passage à la concurrence** : Le client rejoint un concurrent direct (par exemple, quitter Spotify pour Deezer), ce qui constitue un défi frontal pour l'entreprise.

3. **Le churn interne** : Plus subtil, ce phénomène se produit lorsqu'un client migre vers une autre offre au sein de la même entreprise (comme passer d'une carte bancaire classique à une carte premium). Il signale une redéfinition de la fidélité, sans toutefois entraîner une perte totale.

1.1.3 Mesurer le churn : évaluer l'ampleur des pertes

Le taux d'attrition, ou churn rate, se calcule simplement : nombre de clients perdus divisé par le nombre total de clients sur une période donnée. Ce chiffre, bien qu'il varie selon le secteur et les méthodes de calcul, est toujours révélateur de la santé de la base clientèle.

Il est également pertinent de calculer le *revenue churn rate*, qui mesure la perte de revenus sur la même période. Cette analyse plus fine permet de différencier les clients à faible valeur de ceux dont la défection constitue un véritable choc financier. Un churn rate élevé ne signifie pas nécessairement une chute dramatique des revenus, mais peut en être un indicateur précurseur.

Toutefois, il n'existe pas de norme universelle pour le churn. Son taux fluctue selon les secteurs : il dépasse rarement 5 % dans la finance, tandis qu'il avoisine les 20 % dans la téléphonie mobile. C'est une mesure relative mais essentielle.

1.1.4 Maîtriser le churn : les armes du stratège

Pour contenir ce fléau, il est indispensable de se doter d'outils de surveillance performants. Une plateforme de Customer Success permet de suivre chaque client, d'anticiper les défections grâce à des indicateurs comme le score de santé, et de prendre des mesures préventives avant qu'il ne soit trop tard. La force du stratège réside dans la prévention, non dans la réaction.

1.1.5 Le churn : un indicateur clé de la stratégie d'entreprise

Le churn est un reflet de l'efficacité des stratégies mises en place. Pour en tirer pleinement parti, il doit être suivi en continu, notamment après chaque lancement de nouvelles fonctionnalités. Une analyse mensuelle permet de détecter les tendances et de mesurer l'impact des décisions sur la satisfaction client. Ce KPI est un guide précieux pour anticiper les périodes de fragilité de la fidélité, où la concurrence se fait plus menaçante.

1.1.6 Prévenir le churn : l'écoute, la clé de la fidélisation

Un taux de churn élevé est une alerte, un signal d'action. Pour y répondre efficacement, il faut remonter à la source : les clients. Les écouter, comprendre leurs frustrations, et leur accorder une attention soutenue. L'analyse des feedbacks utilisateurs est un levier puissant pour renforcer la fidélité, transformant le mécontentement en loyauté.

1.2 Objectifs

Dans notre projet, nous allons analyser de près l'attrition dans une banque avec un jeu de données en accès libre trouvées sur Kaggle.

Notre objectif est dans un premier temps de déterminer quels facteurs ou combinaison de facteurs influencent l'attrition. Nous utiliserons ensuite ces résultats pour détecter les clients susceptibles de quitter la banque puis proposer un plan d'action automatisé envers ces clients. Cela pourrait se traduire par la mise en place d'une stratégie de la banque pour fidéliser ses clients selon les étapes suivantes :

1/ Identifier les clients concernés : utiliser les données historiques pour comprendre quels indicateurs jouent un rôle dans le départ des clients. A partir de ces données essayer de construire un algorithme fiable pour détecter les clients qui entrent la catégorie de ceux susceptibles de quitter la banque.

2/ Envoyer une alerte au conseiller bancaire du client identifié en proposant des actions commerciales pré-identifiées : à des intervalles réguliers (hebdomadaires par exemple), chaque conseiller bancaire reçoit un listing de ses clients identifiés comme entrant dans une phase de « mécontentement » et donc potentiellement sur le départ. En fonction de la

typologie des facteurs d'insatisfaction, le conseiller bancaire reçoit les actions commerciales à lancer en direction de ces clients sous forme de process ou d'actions (joindre le client par téléphone, proposer un email prérédigé etc.).

3/ Faire un suivi des actions et mesurer leur efficacité : généré un nouveau set de données pour enrichir la stratégie de la banque sur le churn. Tirer avantage des actions engagées pour mieux comprendre les clients et enrichir la connaissance de la banque des ses clients pour augmenter leur satisfaction et ainsi leur fidélisation, notamment à la suite des interactions avec les conseillers bancaires et l'analyse des feedbacks.

2 Gestion des données

2.1 Collecte et pré-traitement

2.1.1 Importation et Examen des Données

Nous avons débuté notre analyse par l'importation et l'exploration minutieuse des données clients que nous avons trouvées sur Kaggle. Cette étape initiale était essentielle pour vérifier la qualité des données, évaluer leur complétude, et s'assurer de leur adéquation pour une analyse approfondie.

- **Données analysées** : Un total de 10 000 enregistrements clients a été utilisé, comprenant des informations démographiques, financières, et comportementales.

- **Objectif** : Identifier les premières tendances et valider la pertinence des variables disponibles, assurant ainsi une base solide pour les étapes suivantes.

L'inspection initiale des données nous a permis de détecter d'éventuelles incohérences ou manques dans les enregistrements, mais également d'acquérir une vue d'ensemble sur la diversité et la distribution des variables disponibles.

2.1.2 Analyse de la Distribution des Scores de Crédit

L'analyse préliminaire a porté sur la répartition des scores de crédit parmi les clients. Cette étape nous a permis de détecter des segments potentiellement à risque et d'éclairer nos efforts futurs.

- **Insight principal** : La distribution des scores de crédit au sein de notre base de clients présente des variations significatives. Cela suggère que certains segments pourraient être plus vulnérables au churn, en particulier ceux avec une situation financière plus fragile.

Cette observation met en lumière l'importance d'examiner les facteurs financiers dans notre stratégie de rétention client.

2.1.3 Identification des Facteurs de Churn

En poursuivant l'analyse, nous avons comparé les scores de crédit entre les clients ayant quitté l'entreprise et ceux restés fidèles. Cette comparaison visait à identifier les caractéristiques distinctives des clients à risque de churn.

- **Résultat clé** : Les clients avec un score de crédit plus bas affichent une tendance plus marquée à quitter l'entreprise. Cette découverte souligne la nécessité d'interventions ciblées pour ces segments spécifiques, potentiellement via des offres personnalisées ou des communications renforcées.

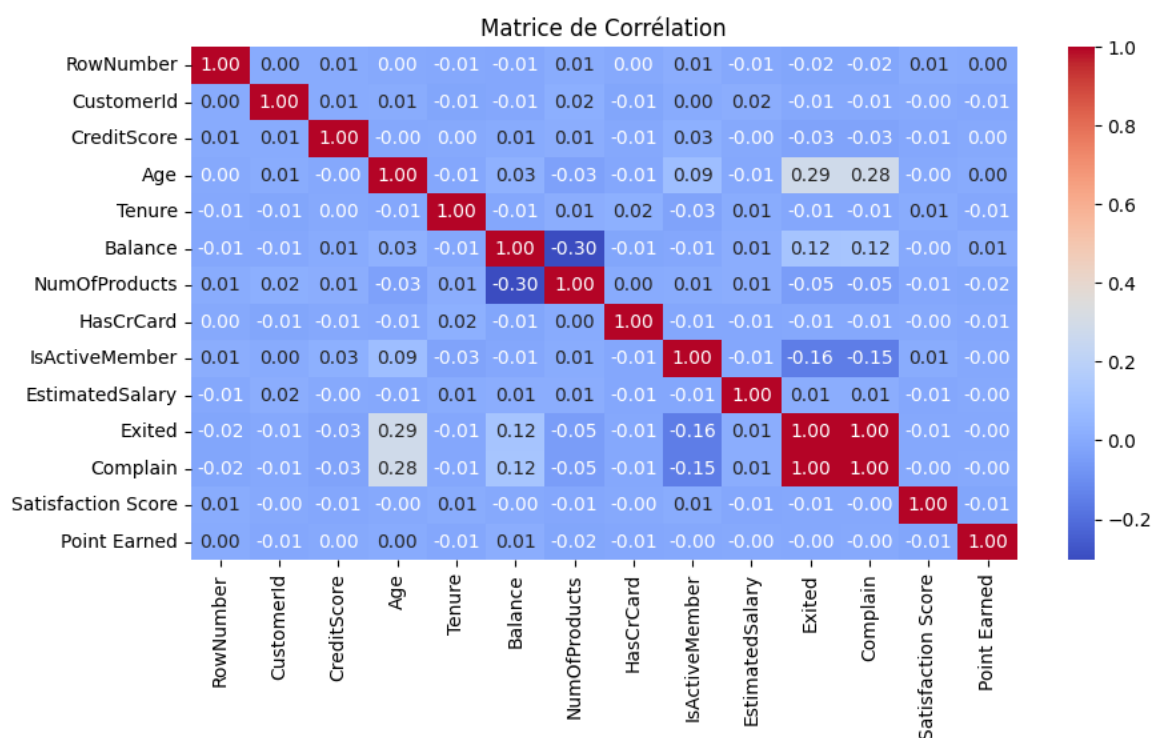
Cette corrélation forte entre le score de crédit et le churn est un indicateur crucial qui orientera nos futures stratégies de rétention.

2.1.4 Analyse des corrélations

Afin d'approfondir notre compréhension des facteurs influençant le churn, nous avons examiné les relations entre différentes variables présentes dans les données à travers une matrice des corrélations.

- **Découverte majeure** : Outre le score de crédit, plusieurs autres variables montrent des corrélations intéressantes avec le churn. Parmi celles-ci, l'âge des clients, le nombre de produits souscrits, et le solde des comptes se distinguent comme des indicateurs potentiels de risque.

Ces insights sont précieux pour hiérarchiser les actions à entreprendre et permettent de concevoir des modèles prédictifs plus précis.



2.1.5 Préparation pour la modélisation

Afin de passer à l'étape suivante, celle de la modélisation, il a été nécessaire de préparer les données. Cette préparation inclut la transformation de variables catégorielles en formats numériques, condition indispensable pour intégrer ces données dans des modèles prédictifs complexes.

- **Prochaines étapes** : Les modèles ainsi développés permettront de prédire de manière plus précise les clients à risque de churn, et de mettre en place des stratégies de rétention personnalisées et efficaces.

2.2 Qualité des données

2.2.1 Préparation des données

Avant de procéder à l'entraînement du modèle, une étape cruciale de préparation des données a été réalisée. Cette préparation visait à garantir que les prédictions soient non seulement fiables mais également précises.

- **Nettoyage des données** : Les colonnes jugées non pertinentes, telles que RowNumber, Complain, CustomerId, et Surname, ont été supprimées, car elles n'apportaient aucune valeur ajoutée à la prédiction du churn. Les Notebook précisent les colonnes supprimées ainsi que les raisons pour lesquelles elles ne sont pas retenues.

- **Encodage des variables catégorielles** : Les variables catégorielles comme Geography, Gender, et Card Type ont été converties en variables numériques à travers un encodage one-hot. Par exemple, la variable Geography a été transformée en plusieurs colonnes (Geography_France, Geography_Germany, etc.), chacune indiquant la présence ou l'absence du client dans une région spécifique.

Cette étape d'encodage a été déterminante pour permettre une exploitation optimale des données dans les modèles de Machine Learning.

2.2.2 Normalisation des données

La normalisation des données constitue une autre étape clé dans la préparation des données, garantissant que chaque variable contribue équitablement à la modélisation.

- **Utilisation de StandardScaler** : La méthode StandardScaler a été appliquée pour normaliser les données, assurant ainsi que chaque caractéristique ait une moyenne de 0 et un écart-type de 1.

La normalisation est particulièrement cruciale pour les algorithmes de Machine Learning qui sont sensibles à l'échelle des données, tels que ceux basés sur la distance. Ce processus permet de renforcer la précision des modèles prédictifs en harmonisant l'impact de chaque variable.

3 Méthodologie

3.1 Choix des algorithmes

3.1.1 Mise en Place et Test des Algorithmes

Après avoir préparé et normalisé les données, la prochaine étape cruciale de notre projet consistait à identifier le modèle de Machine Learning le plus performant pour prédire le churn client. Pour ce faire, nous avons testé plusieurs algorithmes, en nous concentrant sur leur précision et leur capacité à bien généraliser sur des données non vues.

3.1.2 Sélection des caractéristiques

Dans un premier temps, les données préparées ont été soigneusement séparées en deux ensembles : les variables explicatives (X) et la variable cible (y). Les variables explicatives (X) incluent l'ensemble des facteurs susceptibles d'influencer le churn, tels que l'âge, le solde des comptes, le nombre de produits souscrits, et le score de crédit. La variable cible (y) correspond au statut de churn (Exited), indiquant si un client a quitté l'entreprise ou non.

Cette séparation est une étape essentielle pour l'entraînement des modèles, car elle permet de définir clairement ce que le modèle doit prédire (le churn) à partir des caractéristiques observées (les variables explicatives).

3.1.3 Choix du modèle : test et évaluation

Nous avons ensuite procédé à l'évaluation de plusieurs modèles de Machine Learning afin de déterminer lequel offrait les meilleures performances en termes de précision prédictive et de capacité de généralisation. Les modèles testés incluent :

- Régression Logistique
- Forêt d'arbres décisionnels (Random Forest)
- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM)
- Arbre Décisionnel (Decision Tree)

Chaque modèle a été évalué en utilisant une technique de validation croisée, où les données sont divisées en plusieurs sous-ensembles afin de tester la performance du modèle sur des données qu'il n'a pas encore vues. Cette approche nous a permis d'estimer de manière fiable la capacité du modèle à généraliser sur de nouvelles données.

3.1.4 Modèle retenu : arbre décisionnel

Après des tests rigoureux, nous avons retenu Random Forest comme le modèle le plus performant pour notre cas d'usage. Ce choix a été motivé par plusieurs facteurs clés :

- **Précision** : Random Forest a montré une excellente précision prédictive, surpassant d'autres modèles dans notre ensemble de tests.

- **Interprétabilité** : Contrairement à certains modèles plus complexes, comme les SVM ou les réseaux neuronaux, Random Forest offre une grande transparence, permettant de visualiser et d'expliquer facilement les décisions prises par le modèle. Cette interprétabilité est cruciale pour identifier les facteurs spécifiques contribuant au churn.

- **Capacité de Généralisation** : Random Forest a démontré une bonne capacité de généralisation sur des données non vues, ce qui est essentiel pour garantir que les prédictions resteront précises lorsqu'elles seront appliquées à de nouveaux clients.

3.1.5 Conclusion sur le modèle

L'adoption de Random Forest comme modèle principal nous permet d'obtenir des prédictions fiables tout en conservant une compréhension claire des décisions du modèle. Cette combinaison de précision et de transparence est particulièrement importante pour guider les actions de rétention client, en permettant non seulement d'identifier les clients à risque, mais aussi de comprendre les raisons sous-jacentes à leur possible départ. Les prochains développements se concentreront sur le déploiement de ce modèle en production, ainsi que sur l'optimisation continue de ses performances à mesure que de nouvelles données deviendront disponibles.

	Model	Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	0.706333	0.799501	0.706333	0.731953
1	Random Forest	0.862667	0.854999	0.862667	0.856591
2	Gradient Boosting	0.795333	0.840923	0.795333	0.808944
3	Support Vector Machine	0.792000	0.833990	0.792000	0.805147
4	K-Nearest Neighbors	0.713000	0.780964	0.713000	0.734946
5	Decision Tree	0.800667	0.803300	0.800667	0.801935
6	Naive Bayes	0.710667	0.794427	0.710667	0.734974
7	Extra Trees	0.854333	0.845554	0.854333	0.837144
8	xGboost	0.854286	0.845554	0.854333	0.837144
9	xGboost	0.856714	0.845554	0.854333	0.837144

3.2 Approche théorique (justification théorique des choix méthodologiques)

3.2.1 Identification du problème d'Overfitting

Au cours de l'entraînement du modèle d'arbre de décision, nous avons identifié un problème critique : l'Overfitting. Ce phénomène survient lorsque le modèle devient excessivement complexe et commence à "mémoriser" les données d'entraînement au lieu de les généraliser,

c'est-à-dire de capter les tendances sous-jacentes applicables à de nouvelles données. L'Overfitting réduit donc la capacité du modèle à faire des prédictions précises sur des données qu'il n'a pas encore vues.

- **Symptôme** : Le modèle affichait une précision exceptionnellement élevée sur les données d'entraînement, mais sa performance chutait drastiquement lors des tests sur des données non vues. Ce décalage important entre les performances sur l'ensemble d'entraînement et l'ensemble de test indiquait clairement que le modèle était sur ajusté, capturant des détails spécifiques aux données d'entraînement qui ne se généralisaient pas.

3.2.2 Stratégies de gestion de l'Overfitting

Pour pallier ce problème d'Overfitting et améliorer la capacité de généralisation du modèle, nous avons mis en œuvre plusieurs stratégies.

a. Élagage de l'arbre de décision (Pruning)

La première méthode utilisée pour réduire la complexité du modèle a été l'élagage de l'arbre de décision. Cette technique consiste à limiter la profondeur de l'arbre, ce qui force le modèle à prendre en compte des segments de données plus larges plutôt que de se concentrer sur des cas spécifiques et isolés.

- **Objectif** : En restreignant la profondeur de l'arbre, nous avons réduit la complexité du modèle, évitant ainsi qu'il ne surapprenne les détails particuliers des données d'entraînement. Cela permet d'améliorer la robustesse du modèle et de maintenir une meilleure performance sur des ensembles de données non vus.

b. Validation croisée (Cross-Validation)

Enfin, nous avons mis en place la validation croisée pour évaluer la performance du modèle de manière plus rigoureuse. La validation croisée consiste à diviser les données en plusieurs sous-ensembles et à entraîner le modèle sur certaines portions tout en le testant sur d'autres. Ce processus est répété plusieurs fois, chaque sous-ensemble jouant alternativement le rôle d'ensemble d'entraînement et de test.

- **Résultat** : La validation croisée nous a permis d'évaluer la performance du modèle de manière plus fiable, en s'assurant que les résultats étaient cohérents et non biaisés par un ensemble de données spécifique. Cette approche a renforcé notre confiance dans la capacité du modèle à généraliser sur de nouvelles données.

3.2.3 Conclusion

Grâce à ces interventions, élagage de l'arbre et validation croisée, nous avons réussi à atténuer le problème d'Overfitting. Ces ajustements ont permis au modèle de maintenir une bonne performance sur des données non vues, assurant ainsi sa fiabilité en conditions réelles. Ces stratégies renforcent la robustesse du modèle, le rendant apte à prédire efficacement le churn client et à guider les décisions stratégiques en matière de rétention. Les prochaines étapes

consisteront à surveiller continuellement les performances du modèle en production et à effectuer des ajustements si nécessaire, afin de garantir une prédiction précise et cohérente dans le temps.

4 Performance du modèle

4.1 Amélioration du modèle post-Overfitting

Après avoir mis en œuvre des stratégies pour atténuer l'Overfitting, telles que l'élagage de l'arbre de décision, l'utilisation de méthodes d'ensemble comme le Random Forest, et la validation croisée, nous avons obtenu un modèle plus équilibré. Ce modèle révisé a démontré une capacité accrue à prédire le churn avec une plus grande précision sur des données nouvelles, confirmant ainsi l'efficacité des ajustements apportés.

4.2 Résultats améliorés

Les ajustements effectués ont permis d'améliorer significativement la performance du modèle sur l'ensemble de test.

- **Précision accrue** : La précision du modèle sur les données de test a considérablement augmenté, passant de niveaux initialement insatisfaisants à des résultats beaucoup plus robustes. Cela montre que le modèle est désormais mieux équipé pour généraliser les tendances observées dans les données d'entraînement, offrant ainsi des prédictions plus fiables sur de nouvelles données.

Cette amélioration est essentielle pour garantir que les décisions prises sur la base des prédictions du modèle soient pertinentes et applicables à l'ensemble de notre clientèle, et non uniquement aux clients ayant servi à entraîner le modèle.

4.3 Identification des variables influentes

Au-delà des améliorations en termes de précision, l'analyse du modèle révisé a également permis d'identifier les variables ayant le plus grand impact sur les prédictions de churn. Ces insights sont cruciaux pour orienter les stratégies de rétention client.

- **Score de crédit** : Le score de crédit reste l'une des variables les plus influentes, confirmant son rôle critique dans la prédiction du churn. Les clients avec des scores de crédit plus bas sont plus susceptibles de quitter l'entreprise, ce qui justifie des actions ciblées pour améliorer leur satisfaction.

- **Statut d'activité (IsActiveMember)** : Le statut d'activité d'un client (s'il est un membre actif ou non) s'est également révélé être un facteur déterminant. Les membres actifs sont généralement moins susceptibles de cherner, soulignant l'importance d'engager régulièrement les clients pour maintenir leur fidélité.

- **Type de carte de crédit détenue** : Le type de carte de crédit que détient un client joue également un rôle significatif. Certaines cartes semblent associées à un risque de churn plus élevé, ce qui pourrait indiquer des préférences ou des besoins spécifiques non satisfaits par certains produits.

4.4 Conclusion

Grâce aux ajustements apportés pour réduire l'overfitting, nous avons développé un modèle à la fois précis et robuste, capable de faire des prédictions fiables sur le churn client. L'identification des variables les plus influentes nous offre un aperçu précieux des facteurs à surveiller et à améliorer pour renforcer la rétention client. En intégrant ces insights dans nos stratégies d'affaires, nous pouvons anticiper plus efficacement les risques de churn et mettre en place des interventions ciblées pour fidéliser notre clientèle. Les prochaines étapes incluront l'application de ce modèle dans des scénarios réels et l'ajustement continu basé sur les retours du terrain, assurant ainsi un alignement optimal entre les prédictions et les actions entreprises.

5 Perspectives

5.1 Perspectives et recommandations business

Dans le cadre de notre stratégie d'optimisation de la rétention client, nous avons imaginé une application web innovante dédiée aux banques, conçue pour anticiper et gérer efficacement les risques de churn. En s'appuyant sur notre algorithme, cette solution permet de détecter les clients présentant une probabilité élevée de quitter la banque, et offre deux modes d'utilisation flexibles, adaptés aux besoins variés des équipes bancaires.

Tout d'abord, l'application propose un mode interactif où l'utilisateur peut saisir directement les informations relatives à un client spécifique. Grâce à notre algorithme, l'application génère instantanément un diagnostic indiquant si ce client est à risque de churn, permettant ainsi d'agir rapidement et de façon ciblée. Ce mode d'utilisation est idéal pour les conseillers clientèle qui cherchent à gérer des interactions individuelles de manière proactive et personnalisée.

Ensuite, pour une analyse à plus grande échelle, l'application permet de traiter une base de données entière de nouveaux clients au format Excel. L'algorithme extrait les profils les plus à risque, fournissant ainsi à l'équipe bancaire une vue d'ensemble des clients nécessitant une attention prioritaire. Cette fonctionnalité permet d'anticiper les risques à une échelle beaucoup plus large et de maximiser les efforts de rétention.

Mais nous ne nous arrêtons pas là. Pour apporter une valeur ajoutée supplémentaire, nous avons intégré une fonctionnalité qui propose des stratégies d'intervention sur mesure en fonction des causes potentielles de churn identifiées. Par exemple, pour les clients rencontrant des problèmes de communication avec la banque, l'outil pourrait proposer un modèle d'e-mail ou un script pour un appel téléphonique, renforçant ainsi le lien avec le client. De même, pour ceux dont le churn est lié à un manque d'options ou à une insatisfaction concernant les produits, l'application suggère des solutions adaptées, telles que de nouveaux produits ou services qui pourraient répondre à leurs besoins spécifiques. L'objectif ultime est de créer une interaction significative avec le client, tout en profitant de chaque échange pour affiner les données sur ses préférences et comportements.

Enfin, dans notre démarche nous proposons d'inclure plus tard un suivi rigoureux des performances de l'algorithme. En analysant les statistiques, nous visons à affiner continuellement notre modèle pour garantir des résultats encore plus précis. Nous mesurons notamment combien de clients à risque ont été fidélisés, combien sont partis malgré l'intervention, et combien de clients non identifiés comme à risque ont quitté la banque. Cette boucle de rétroaction permettra non seulement d'améliorer l'efficacité de l'algorithme, mais également de renforcer l'impact global de la stratégie de rétention client.

Avec cette application, la banque dispose d'un puissant levier pour non seulement prévenir les départs, mais aussi transformer chaque interaction à risque en une opportunité d'engagement client. A ce stade nous n'avons pas développé l'outil génératif mais nous recommanderions au client d'aller vers une telle solution compte tenu des apports qu'elle aurait.

6 MLOps

6.1 *Déploiement*

Après avoir finalisé le développement de l'algorithme et évalué ses performances, il faut déployer l'application ainsi que le modèle dans un environnement de production stable et scalable (cloud par exemple).

6.2 *Surveillance et maintenance*

Théoriquement, une fois l'algorithme déployé en production, la surveillance et la maintenance constitueraient des étapes clés pour garantir son bon fonctionnement à long terme. Nous présentons ci-dessous une description de cette démarche qui ne se limite pas à la simple correction des erreurs, mais comprend également la surveillance continue de la performance des modèles, leur mise à jour, et l'intégration de nouvelles données sans perte de qualité. Nous allons détailler la stratégie à mettre en place pour assurer ces aspects, avec un accent particulier sur la mise en œuvre rigoureuse des tests de non-régression afin de détecter toute dégradation de performance au fil du temps.

L'objectif principal de la stratégie de maintenance est double :

1. Assurer la stabilité de la performance des algorithmes à travers des tests réguliers qui comparent les résultats actuels à des performances de référence (baseline).
2. Faciliter l'amélioration continue des modèles en évaluant l'impact des mises à jour et des révisions du code ou des données sur la qualité des prédictions.

6.2.1 Mise en place d'une métrique de performance de référence

La première étape pour garantir une gestion optimale de la maintenance des algorithmes est de définir des métriques de performance claires qui serviront de référence. Les métriques à utiliser varieront selon la nature des modèles (classification, régression, clustering, etc.) et les objectifs du projet.

Ces mesures serviront de baseline à un instant T, c'est-à-dire lors de la première validation du modèle avant son déploiement en production. La baseline sera calculée sur un jeu de test représentatif et sera stockée afin de permettre des comparaisons ultérieures.

6.2.2 Surveillance continue et tests de non-régression

a. Déclenchement automatique des tests

Pour éviter toute dérive dans la performance des algorithmes, il est crucial de mettre en place un pipeline automatisé permettant de tester régulièrement les modèles en production. Ces tests incluent :

- Tests périodiques (quotidiens, hebdomadaires) pour surveiller les performances en continu.
- Tests déclenchés par des événements, notamment lors de la réception de nouvelles données, des mises à jour du modèle ou des modifications du code (ajustements hyperparamétriques, ajout de nouvelles fonctionnalités).

Ces tests comparent les nouvelles performances à celles de la baseline, permettant ainsi d'identifier toute dégradation.

b. Utilisation de jeux de test distincts

Afin de garantir que les tests reflètent fidèlement les conditions en production, plusieurs types de jeux de données seront utilisés :

- Jeu de test initial : conservé depuis la première validation du modèle, il sert de référence absolue.
- Jeux de données simulant les nouveaux flux de données : intégrés pour tester la capacité du modèle à s'adapter à des distributions changeantes (phénomène de drift).
- Jeux de test adverses : créés pour évaluer la robustesse du modèle face à des entrées déviantes ou bruitées.

Ces différents jeux permettent d'assurer une évaluation exhaustive du modèle, en testant ses performances non seulement sur des cas typiques mais aussi sur des cas limites.

6.2.3 Gestion du concept drift

L'un des principaux défis en matière de maintenance des modèles MACHINE LEARNING est le concept drift : lorsque la relation entre les variables explicatives et la variable cible évolue dans le temps.

Pour détecter ces changements nous proposons :

- Des tests statistiques sur les distributions des nouvelles données (Kolmogorov-Smirnov, test de chi-deux) seront exécutés régulièrement pour comparer la distribution des nouvelles données avec les anciennes.
- Des outils de surveillance active du drift comme les techniques de détection basées sur des fenêtres glissantes seront déployés pour alerter lorsqu'une modification significative du comportement des données ou des performances est détectée.

6.2.4 Versionnement des modèles et rollback

Chaque version du modèle déployée en production sera soigneusement documentée et versionnée à l'aide d'outils comme MACHINE LEARNING flow ou DVC (Data Version Control). Cette pratique permet de :

- Tracer les modifications effectuées dans les données, les hyperparamètres ou les fonctionnalités des modèles.
- Faciliter le rollback à une version antérieure en cas de régression inattendue de la performance après une mise à jour.
- Assurer la reproductibilité des expériences pour permettre une analyse approfondie des modifications apportées.

6.2.5 Enrichissement et réentraînement des modèles

Afin de maintenir et améliorer les performances à long terme, il est nécessaire de planifier des sessions de réentraînement régulier des modèles, notamment lorsque de nouveaux jeux de données sont disponibles. Cette approche permet d'adapter les modèles aux évolutions du domaine d'application, aux nouveaux patterns des données, et aux éventuelles modifications des distributions sous-jacentes. Les tests de non-régression seront systématiquement appliqués avant tout réentraînement afin de s'assurer que les nouvelles données n'introduisent pas de biais ou de régressions.

Un processus d'A/B testing sera également intégré pour mesurer l'impact des nouveaux modèles avant tout déploiement à grande échelle, en comparant les résultats des anciens et nouveaux modèles sur une partie du trafic réel.

6.2.6 Outils et Infrastructure

Pour garantir l'efficacité de ce processus, une infrastructure adaptée sera mise en place :

- CI/CD pour Machine Learning : un pipeline de Continuous Integration et Continuous Deployment sera utilisé afin de garantir que chaque nouvelle version du modèle est testée de manière automatique avant déploiement.

- Monitoring en temps réel : des solutions de surveillance (comme Prometheus ou Grafana) seront utilisées pour suivre en temps réel les métriques de performance des modèles en production.
- Logging et alertes : chaque prédiction, ainsi que ses métriques associées, seront loguées et des systèmes d'alerte seront configurés pour notifier l'équipe en cas de dégradation de performance significative.