REPUBLIQUE D'HAITI

AKADEMI

RAPPORT D'ANALYSE DES GENRES DE FILMS LES PLUS PROMETEURS

PRESENTÉ PAR : DEBREUS Monitès

PROFESSEURS : JEROME Wedter, LAGUERRE Geovany

Date : 24 Aout 2025

Information personelle

Email :monites.debreus@student.ueh.edu.ht

Profil linkedin : www.linkedin.com/in/monites-debreus-7b4b7a1ab

Aperçu du travail

Ce travail constitue un aperçu analytique réalisé dans le cadre d'un projet visant à aider une compagnie de télécommunications à mieux comprendre et anticiper le phénomène de résiliation de ses clients (churn). Dans un contexte de forte concurrence et de volatilité des abonnés, l'entreprise cherche à s'appuyer sur une étude fondée sur des données concrètes afin d'améliorer sa capacité de rétention.

Pour ce faire, nous exploitons un jeu de données regroupant diverses informations sur les abonnés (caractéristiques contractuelles, utilisation des services, variables démographiques et historiques de facturation). L'analyse repose sur des techniques de modélisation statistique et d'apprentissage automatique permettant de prédire la probabilité qu'un client se désabonne.

Les résultats de ce travail offrent un cadre d'aide à la décision : ils permettent non seulement d'identifier les facteurs les plus déterminants du churn, mais aussi de construire des modèles prédictifs capables de détecter les clients à risque. Ces outils fourniront à l'entreprise des leviers stratégiques pour orienter ses campagnes de fidélisation et maximiser la valeur de sa clientèle.

NB : pour une meilleure compréhension du rapport :

- ✓ churner : client résilié
- ✓ non-churner : client non-resilié
- ✓ churn : résiliation

Table des matières

LISTE DES FIGURES

1 Introduction

Le secteur des télécommunications connaît une dynamique intense, marquée par une concurrence accrue, la diversification des services, et une exigence toujours plus grande des clients en matière de qualité et de réactivité. En 2025, le marché des télécoms continue de croître, avec une adoption massive des services mobiles, de l'internet et des forfaits internationaux, ce qui place les opérateurs devant le défi majeur de fidéliser leurs abonnés.

Dans ce contexte, la résiliation des abonnements, ou churn, représente une menace directe pour la rentabilité et la pérennité des compagnies télécoms comme SyriaTel. Comprendre le comportement des clients, identifier les facteurs favorisant la résiliation, et anticiper les départs deviennent essentiels pour orienter les stratégies marketing et les interventions du service client.

La mission de ce travail est donc de prévoir les abonnés susceptibles de résilier leur contrat en s'appuyant sur un jeu de données riche comprenant des informations sur les caractéristiques des abonnés, l'usage des services, les plans souscrits, la facturation, et le support client. L'objectif est de transformer ces analyses en actions concrètes pour réduire le churn, améliorer la satisfaction client et renforcer la position concurrentielle de SyriaTel .

1.1 Objectifs

Les objectifs de ce travail sont classes en deux catégories : objectif principal et objectifs spécifiques

Objectif principal :

Ce travail est réalisé dans le but de fournir à l'entreprise une compréhension approfondie des facteurs influençant le désabonnement (churn) de ses clients, afin de développer des stratégies prédictives et opérationnelles permettant de réduire le taux de perte de clientèle et d'améliorer la fidélisation.

Objectifs spécifiques :

- D'analyser la distribution des variables clients : caractéristiques démographiques, abonnements, habitudes d'utilisation) et leur impact sur le churn.
- D'examiner le rôle des variables catégorielles et continues dans la prédiction du désabonnement, en mettant en évidence les facteurs les plus déterminants.
- D'appliquer et comparer différents modèles de machine learning (régression logistique, arbres de décision, forêts aléatoires, etc.) pour évaluer leurs performances dans la détection des clients à risque.
- D'améliorer la qualité des prédictions en recourant à des techniques de rééquilibrage des classes et d'optimisation des hyperparamètres
- De proposer un cadre décisionnel basé sur les résultats obtenus, afin d'orienter l'entreprise dans la mise en place de politiques ciblées de rétention et de fidélisation.

2 Public cible

Ce travail s'adresse avant tout à la direction de SyriaTel, notamment aux responsables marketing, service client et stratégie, qui sont chargés de prendre des décisions pour réduire le churn et améliorer la fidélisation des clients. Il vise également les analystes et responsables opérationnels qui seront amenés à mettre en œuvre des actions ciblées sur les abonnés à risque. Ce rapport doit servir de base pour orienter les stratégies de rétention, en identifiant les clients les plus susceptibles de résilier et les facteurs influençant leur comportement.

3 Compréhension des données disponibles

Pour réaliser cette analyse, nous disposons d'un jeu de données de 3 333 clients de SyriaTel, regroupant un ensemble riche de variables décrivant leur profil, leur utilisation des services et leur comportement vis-à-vis de l'entreprise. Ces informations sont essentielles pour comprendre les facteurs influençant le churn et orienter les décisions marketing et de fidélisation.

Particulièrement, les variables incluent des informations sur le type d'abonnement (international plan, voice mail plan), l'activité téléphonique (total day/eve/night/intl minutes, calls, charges), ainsi que l'interaction avec le service client (customer service calls). Ces données permettent une exploration détaillée des comportements des clients, fournissant un fondement solide pour identifier les caractéristiques des abonnés à risque et prédire leur résiliation.

En complément, d'autres variables telles que l'état de résidence ou le code régional peuvent apporter un contexte supplémentaire pour l'analyse, permettant de détecter des tendances géographiques ou démographiques liées au churn. L'ensemble de ces informations servira à entraîner des modèles de classification et à fournir des recommandations pratiques pour réduire la perte de clients et améliorer la fidélisation

4 Méthodologie

Cette section comprend les matériels, logiciels et extension qu'on a utilisé pour la réalisation de ce travail ainsi que la méthode suivie pour parvenir aux résultats escomptés.

4.1 Matériels utilisés

- Ordinateur portatif (hp, core i5, 7^{e} generation): principal support sur lequel on a installé tous les logiciels et extensions nécessaires pour faire l'analyse
- Git (version 2.49.0), Anaconda (version 2024.10-1), Jupyter notebook :respectivement des logiciels et environnement interactif qui permet

d'exécuter du code dans des notebooks organisés en cellules permettant de clean les données, de faire des analyses et des graphiques.

- Chatbot : pour le dépannage des erreurs et des conseils d'écritures de codes
- Les supports du cours pour la révision de l'utilité de certaines fonctions et méthodes en python.
- Word (Ms 365) : pour la rédaction du rapport final.

4.2 Méthode :

Dans cette partie, on présente l'ensemble des démarches suivies pour la réalisation du travail.

4.2.1. Lecture du fichier

Pour la réalisation de ce travail, nous avons utilisé le fichier client de SyriaTel telechargé sur Kaggle, contenant 3333 observations et 21 variables (démographiques, d'usage et de service). Ce fichier a été importé dans un environnement Jupyter Notebook à l'aide de la bibliothèque Pandas afin d'inspecter son contenu et identifier les variables d'intérêt.

Les colonnes disponibles incluent notamment :

- Des données générales : state, area code, phone number, account length.
- Des informations sur les abonnements : international plan, voice mail plan.
- Des données d'usage : total day minutes, total eve minutes, total night minutes, total intl minutes, ainsi que le nombre d'appels (total day calls, total eve calls, etc.).
- Des variables financières associées : total day charge, total eve charge, total night charge, total intl charge.
- Des données de service : customer service calls.
- La variable cible : churn (booléenne indiquant si le client a quitté l'opérateur ou non).

Un aperçu statistique initial des variables a été effectué grâce aux fonctions : .shape(), .info() , .describe() et .value_counts()… afin de comprendre la distribution des données et vérifier l'équilibre des classes dans la variable churn (≈ 14 % des clients ont quitté l'opérateur).

4.2.2. Nettoyage des données

Un processus de préparation a été appliqué pour rendre le dataset exploitable :

- Vérification des valeurs manquantes : aucune donnée manquante n'a été détectée.
- Transformation des variables catégorielles (international plan, voice mail plan, state) en variables numériques via encodage binaire ou One-Hot Encoding.
- Suppression des variables non pertinentes pour la prédiction, telles que phone number, qui n'apporte aucune information discriminante.
- Standardisation des variables continues afin d'améliorer la convergence des modèles basés sur des distances.
- Vérification de la distribution de la variable cible : constat d'un fort déséquilibre (clients churn ~14 %, non-churn ~86 %).

4.2.3. Analyse exploratoire et visualisations

Des visualisations et statistiques descriptives ont été menées pour mieux comprendre le comportement des clients :

- Analyse de la distribution de la variable churn
- Analyse de la distribution des minutes (day, evening, night, international) et leur lien avec le churn.
- Étude de l'impact du service client sur le churn (corrélation entre customer service calls et churn).
- Heatmap de corrélations pour identifier les variables fortement corrélées entre elles et avec la variable cible.
- Comparaison des proportions de churn en fonction des plans (international plan, voice mail plan).

Ces analyses ont permis d'identifier des patterns discriminants : par exemple, les clients ayant un international plan présentent un taux de churn significativement plus élevé.

4.2.4. Modélisation et amélioration des performances

Plusieurs modèles de machine learning ont été testés :

- Régression Logistique : modèle de base permettant une première estimation.
- Arbre de Décision : pour capturer les relations non linéaires.
- Random Forest: pour améliorer la stabilité et la performance en réduisant le surapprentissage.

Les modèles ont été évalués avec les métriques suivantes :

- Accuracy (précision globale).
- Recall et F1-score pour la classe churn (True), considérés comme prioritaires, car l'objectif est de détecter les clients à risque.

Des techniques d'optimisation des hyperparamètres ont été appliquées, notamment sur :

4.2.5. Résultats et Interprétation

L'interprétation se fait étape par étape en fonction de chaque résultat obtenu pour mieux expliquer les modèles.

5 Résultat et Analyse

5.1 Analyse globale du jeu de données:

L'analyse des jeux de données a permis d'observer qu'en moyenne, les clients ont des comptes qui ont en moyenne 101 jours. Ces clients sont concentrés dans des states qui utilisent les area code (415 et 510).

Les clients passent en moyenne 100 appels au cours de la journée tout comme au cours de la nuit d'une durée moyenne de 180 minutes en appels de jour et 201 minutes en appels du soir. L'usage international reste beaucoup plus limité avec seulement près de 5 appels pour une durée moyenne de 10 minutes.

La variable customer service calls montre que la majorité des clients contactent peu le service client (médiane = 1 appel), mais certains vont jusqu'à 9, ce qui pourrait signaler des problèmes de satisfaction ou de fidélisation.

5.2 Distribution des cas de résiliations (focus sur la variabe cible)

Selon la distribution des cas de resiliations, l'entreprise enregistre un taux de 14.5% de résiliations, ce qui est inquiétant pour une telle entreprise par rapport à la concurrence sur le marché.

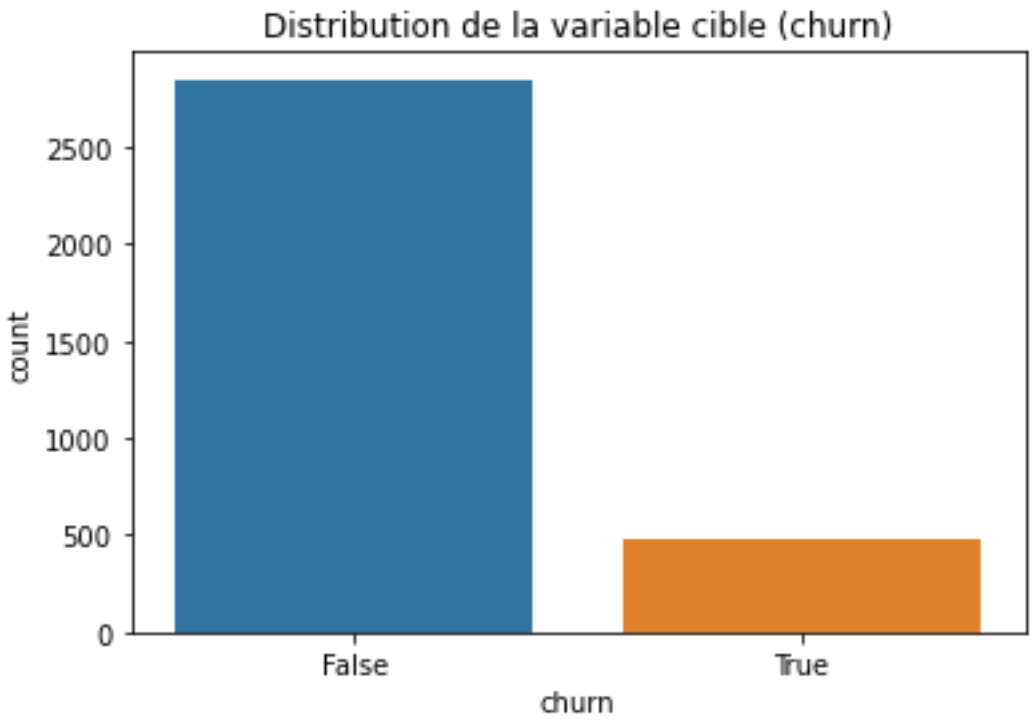

Figure 1: Distribution des cas de résiliations

Toutefois les cas de resiliations peuvent etre influencées par plusieurs facteurs :

- Churn selon le nombre d'appels au service client

Les resultats de cette analyse on montré que proportionnellement la majorité des clients qui appellent 0, 1, 2 ou 3 fois ne résilient pas leur contrat, ce qui est différent pour les clients qui appellent plus de 4 fois. Pour ces clients, le taux de désabonnement devient égal ou même supérieur à celui des clients qui restent. Ce résultat suggère que des problèmes récurrents ou non résolus, forçant les clients à multiplier les appels, sont un indicateur d'une insatisfaction qui influence leur départ.

Le graphique suivant illustre le cas de résiliation des client en fonction du nombre d'appels au service client.

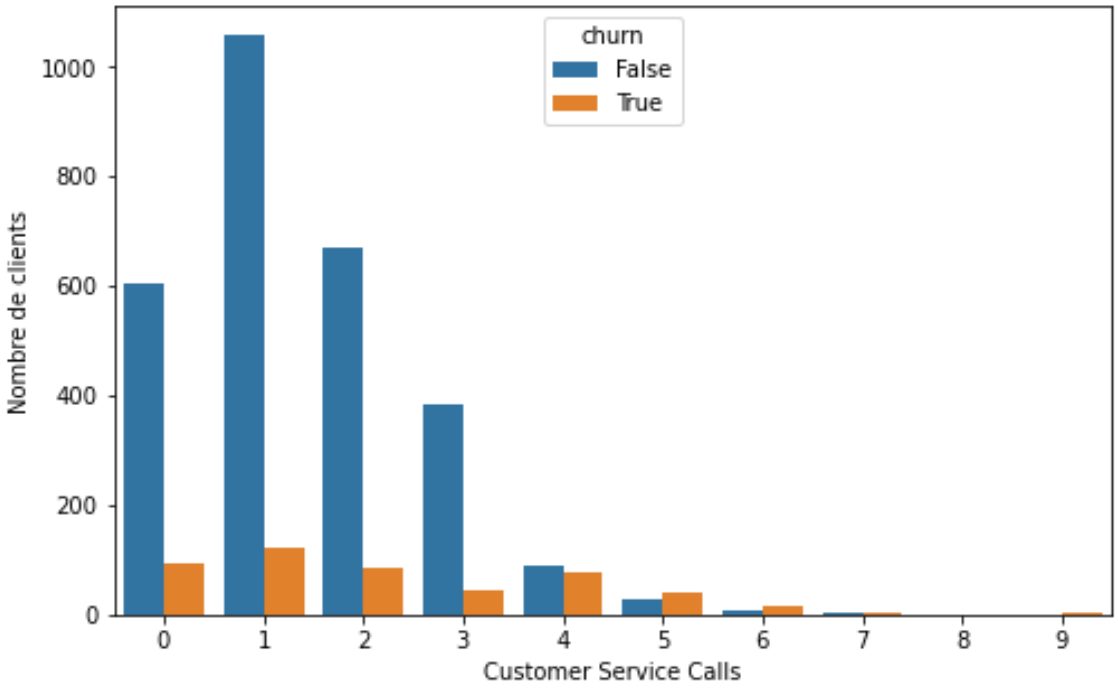

Figure 2: Résiliation selon le nombre d'appels au service client

L'analyse faite sur les cas de résiliation en fonction de la durée des appels de jour et de nuit montre que les clients qui résilient leur contrat ont, en moyenne, une durée d'appels plus longue que les clients qui restent.

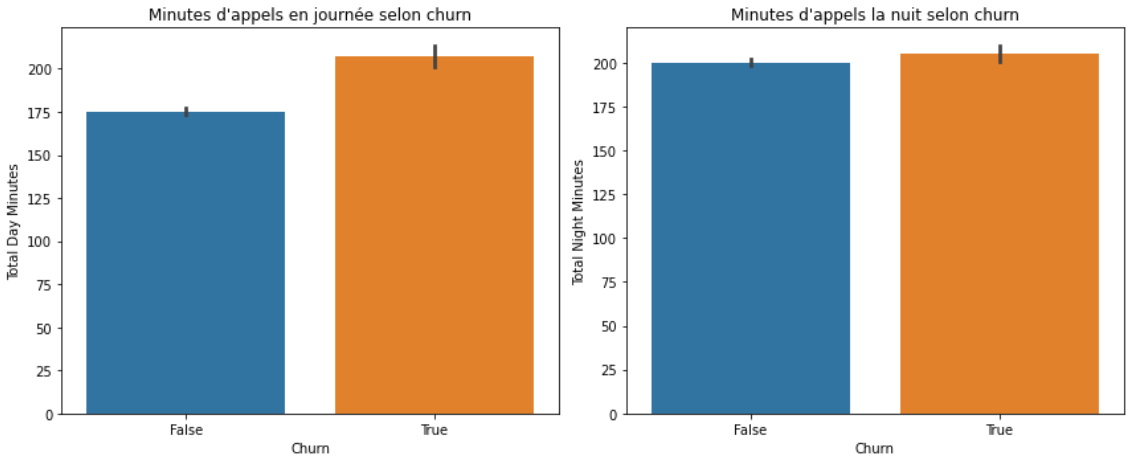

Figure 3:Churn en fonction de la durée des appels

Cela indique en quelque sorte que les clients qui consomment le plus de minutes de communication sont proportionnellement plus susceptibles d'être insatisfaits par les tarifs, la qualité du service, ou qu'ils sont des utilisateurs très actifs qui ont des besoins spécifiques non satisfaits, ce qui conduit finalement à leur départ.

- Churn en fonction de plan international et message vocal

Les analyses montrent que le fait d'avoir un plan international semble être un facteur de risque significatif pour la résiliation. Les clients ayant ce plan sont proportionnellement beaucoup plus susceptibles de partir alors que la messagerie vocale semble être un facteur protecteur contre la résiliation. Les clients qui ont souscrit à ce service sont beaucoup moins susceptibles de résilier leur contrat.

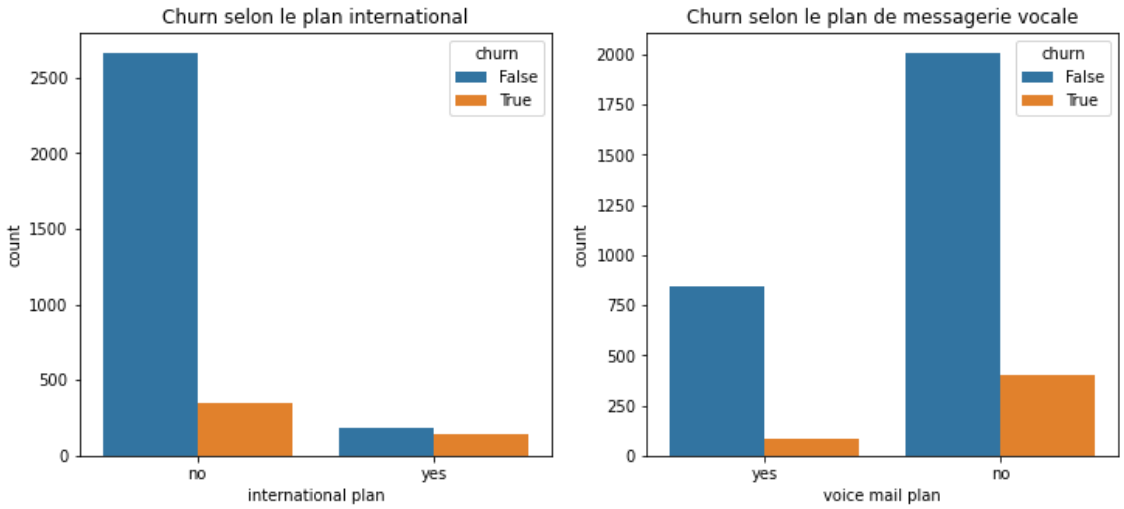

Figure 4: Churn en fonction de plan international et message vocal

Etant donné que les clients ayant ce plan sont proportionnellement beaucoup plus susceptibles de partir, cela indiquerait des problèmes liés à la tarification, à la qualité du service, ou à une offre concurrente plus attrayante pour l'utilisation internationale.

5.3 Corrélation des variables Churn

L'analyse des corrélations entre les variables du jeu de données et la variable cible Churn montre que certaines caractéristiques des clients sont plus fortement associées à la probabilité qu'ils résilient leur abonnement.

Les variables les plus corrélées positivement avec le churn sont : customer service calls (0,209) et total day minutes/total day charge (≈0,205), indiquant que les clients passant plus d'appels au service client ou consommant beaucoup de minutes en journée ont plus de chances de quitter le service comme on l'a montré plus haut.

Les variables liées aux appels en soirée (total eve minutes et total eve charge) et aux appels internationaux (total intl minutes et total intl charge) montrent une corrélation plus faible mais toujours positive, tandis que certaines variables présentent des corrélations négatives, comme total intl calls (-0,053) et number vmail messages (-0,090), suggérant que davantage de messages vocaux ou d'appels internationaux sont légèrement associés à une moindre probabilité de churn. Enfin, account length et area code ont une corrélation

très faible, ce qui indique qu'elles ont peu d'impact direct sur le comportement de résiliation dans ce jeu de données. Globalement, ces corrélations confirment que les comportements d'utilisation et les interactions avec le service client sont les facteurs les plus déterminants pour prédire le Churn (résiliation).

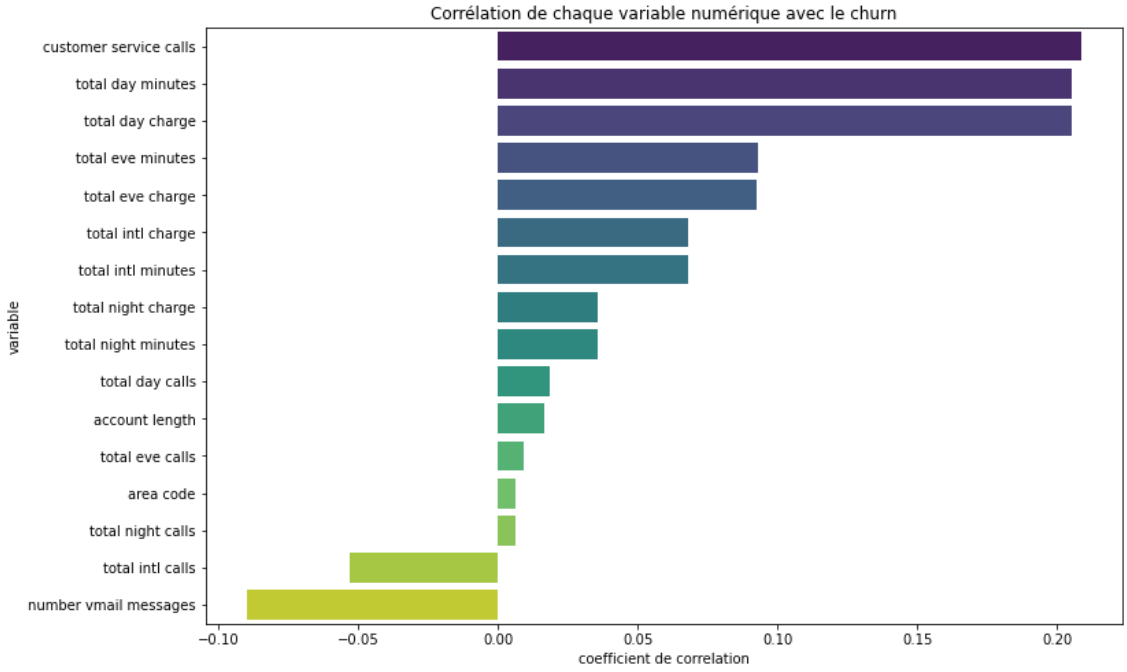

Figure 5:Corrélation des variables numériques avec Churn

5.4 Modélisation

La modélisation a été utilisée pour prédire le comportement des clients vis-à-vis du churn et identifier les facteurs les plus influents. Elle transforme les données brutes en insights exploitables, permettant de détecter les clients à risque et d'anticiper leur résiliation. Plusieurs techniques ont été appliquées, notamment la régression logistique et l'arbre de décision, ces techniques présentent les résultats suivants :

- **Regression Logistique**

Le modèle de régression logistique a été entraîné sur les données initiales, sans rééquilibrage des classes ni réglage des hyperparamètres. Les performances obtenues sont résumées ci-dessous :

- Accuracy globale : 0.84
- Précision (Precision) : 0.85 pour les clients non résiliés, 0.64 pour les clients résiliés ;
- Rappel (Recall) : 0.98 pour les clients non résiliés, 0.19 pour les clients churners ;
- F1-score : 0.91 pour les clients non churners, 0.30 pour les clients churners ;

De ce fait, le modèle identifie très bien les clients qui ne résilient pas, mais il détecte seulement 19 % des clients qui sont résiliés, ce qui montre une faible sensibilité pour la classe minoritaire. Ce résultat met en évidence le déséquilibre du jeu de données et la nécessité de méthodes de rééquilibrage ou de modèles plus adaptés pour mieux prédire le churn.

- **Arbre de décision**

Le modèle d'arbre de décision a été entraîné sur les données initiales, sans rééquilibrage des classes ni réglage des hyperparamètres. Les performances obtenues sont les suivantes :

- Accuracy globale : 0.89 ;
- Précision (Precision) : 0.89 pour les clients non churners, 0.90 pour les clients churners ;
- Rappel (Recall) : 0.99 pour les clients non churners, 0.45 pour les clients churners ;
- F1-score : 0.94 pour les clients non churners, 0.60 pour les clients churners.

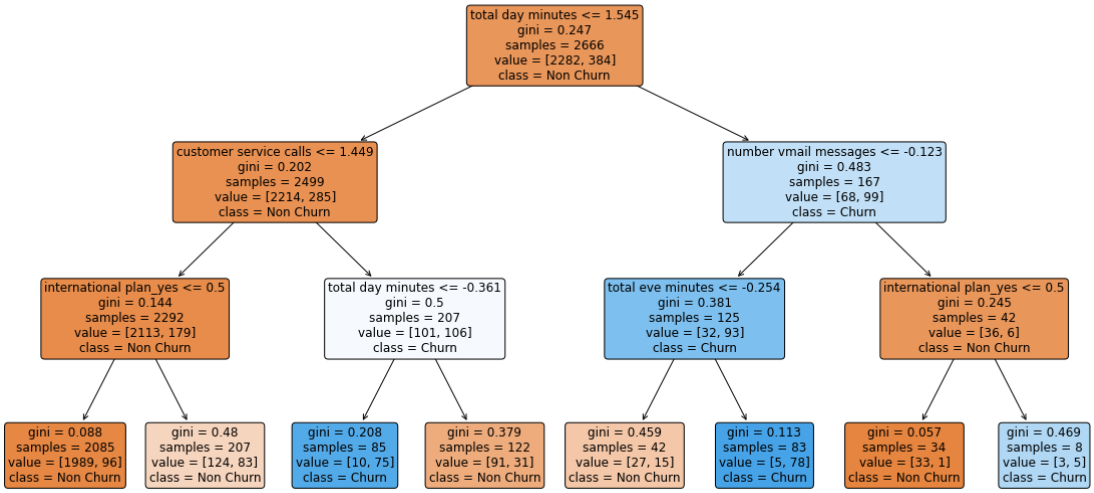

Figure 6: Arbre décisionnel

Selon ce modèle, total day minutes est la variable la plus discriminante pour décrire la résiliation des clients. Les clients qui consomment peu sont les plus fidèles, cependant même si le client consomme peu, avec un taux d'appels élevé au service client, le risque de résiliation augmente.

L'arbre de décision améliore légèrement l'identification des clients churners par rapport à la régression logistique (F1-score passant de 0.30 à 0.60), mais il reste limité pour la classe minoritaire, détectant seulement un tiers des clients churners.

5.5 Comparaison des deux modèles

L'Arbre de Décision est le modèle le plus performant des deux. Il surpasse la Régression Logistique sur toutes les métriques. Cependant, malgré sa haute précision pour prédire

les désabonnements (Precision), sa capacité à détecter la majorité des clients qui résilient (Recall) reste faible, tout comme celle de la Régression Logistique.

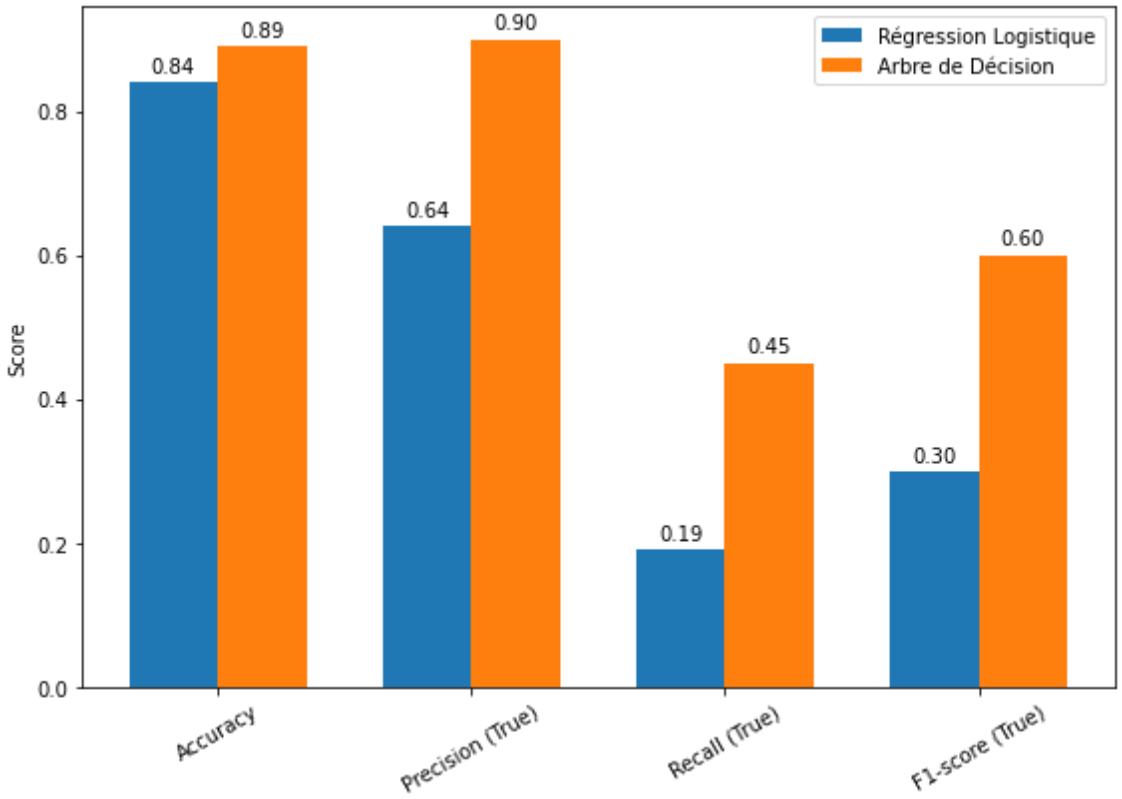

Figure 7: Comparaison des deux modèles

5.6 Amélioration des modèles

L'amélioration des modèles est faite pour corriger le problème de déséquilibre des classes.

- Régression linéaire

Le rééquilibrage des classes a permis d'améliorer considérablement le rappel pour la classe minoritaire (clients churners), passant de 0.19 à 0.71, ce qui signifie que le modèle détecte maintenant 71% des clients susceptibles de churner. Cependant, cette amélioration du rappel a un coût : le score de précision global et l'accuracy ont légèrement

diminué. Ce compromis est typique dans les problèmes déséquilibrés et montre que le modèle est désormais plus sensible aux clients à risque, ce qui est l'objectif principal de cette analyse.

- Arbre de decision

Le modèle amélioré réussit mieux à détecter les clients à risque de churn (classe minoritaire) grâce au rééquilibrage et aux hyperparamètres ajustés, mais il reste un compromis entre précision et rappel pour cette classe. C'est typique pour des problèmes déséquilibrés.

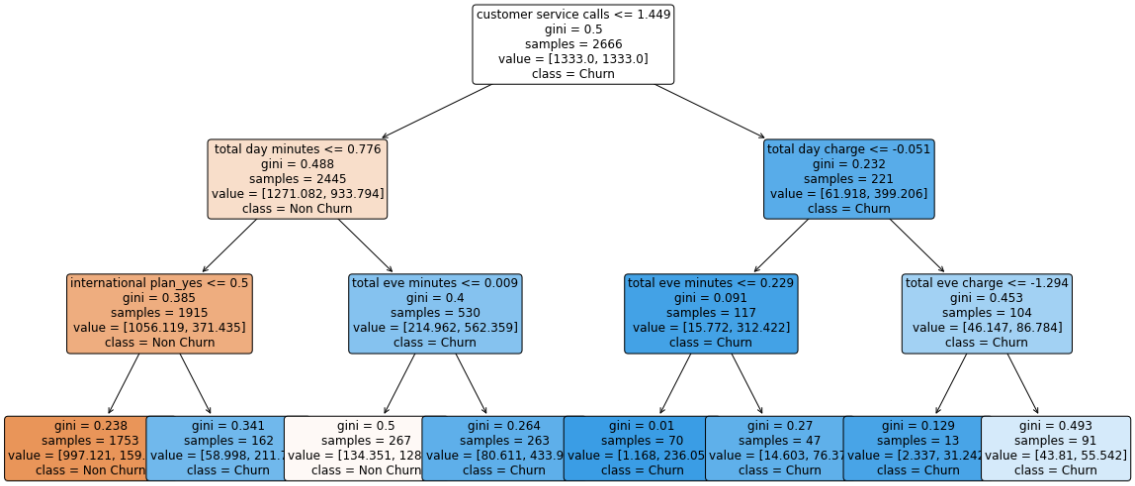

Figure 8: Arbre décisionnel No 2

Cette fois-ci, l'arbre met beaucoup plus d'accent sur le nombre d'appels au service client. Un nombre élevé d'appels au service client place immédiatement un client dans la catégorie à risque. De plus, une forte consommation en journée (total day minutes) est un facteur important qui augmente considérablement la probabilité de résiliation.

5.7 Random Forest

Les résultats du modèle Random Forest montrent une performance globale très satisfaisante, avec une précision (accuracy) de 93 %, ce qui indique que la majorité des

prédictions sont correctes. Pour la classe False (clients qui ne résilient pas), les métriques sont excellentes : une précision de 96 % et un rappel de 96 %, ce qui signifie que le modèle identifie presque parfaitement les clients qui restent. En revanche, pour la classe True (clients qui résilient), les performances sont légèrement moins élevées : la précision est de 80 % et le rappel de 81 %, indiquant que le modèle détecte bien une majorité des churners. Le score macro moyen (0.88) confirme cet écart entre classes, montrant que la classe minoritaire reste plus difficile à modéliser malgré de bons résultats. Globalement, le Random Forest parvient donc à bien équilibrer la détection des deux catégories, mais des améliorations pourraient encore être envisagées pour optimiser la détection des clients en churn.

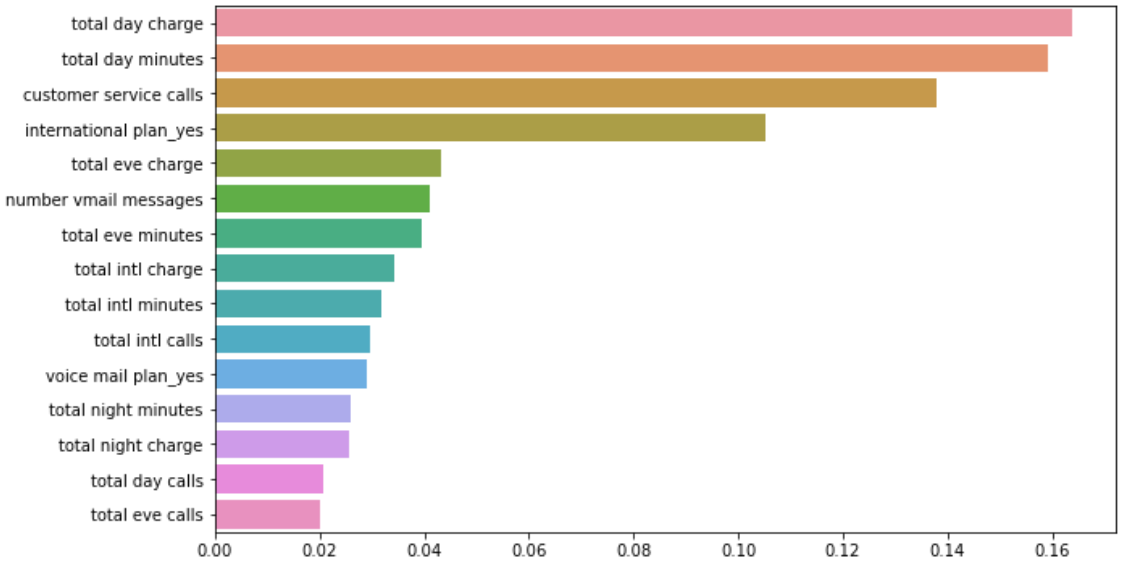

Figure 9: Les variables qui expliquent mieux Random Forest

Selon ce graphique les coûts et la durée des appels, surtout en journée, ainsi que les interactions avec le service client, sont les meilleurs indicateurs du churn.

Un modèle Random Forest accorde plus d'importance à ces aspects pour prédire si un client va partir.

5.8 Matrice de confusion

La matrice de confusion montre en détail la répartition des prédictions correctes et erronées du modèle Random Forest. Sur les 568 clients qui ne résilient pas (classe False), le modèle en a correctement identifié 550, tandis qu'il en a mal classé 18 comme churners. Cela signifie qu'il fait très peu d'erreurs lorsqu'il s'agit de reconnaître les clients fidèles. En ce qui concerne les 99 clients churners (classe True), le modèle en a correctement détecté 79, mais il en a 20 qui ont été prédits à tort comme non-churners. Ce dernier cas est plus problématique, car ce sont des clients à risque qui passent inaperçus, ce qui peut avoir un coût pour l'entreprise. Globalement, la matrice confirme que le modèle est très performant pour identifier les non-churners, mais qu'il reste des progrès à faire pour améliorer la détection des churners, qui constituent pourtant la catégorie la plus critique pour la stratégie de fidélisation..

6 Conclusions et recommandations

Ce travail d'analyse met en évidence plusieurs dynamiques clés du comportement de résiliation des clients. À partir des données d'usage (durée des appels, nombre d'appels internationaux, appels au service client, messagerie vocale, etc.), combinées aux informations de résiliation (churn) et aux modèles de prédiction (régression logistique, arbre de décision, Random Forest), nous avons pu aboutir aux conclusions suivantes :

L'exploration descriptive a montré que la résiliation concerne environ 14,5 % des clients, un taux relativement préoccupant pour le secteur. Parmi les variables explicatives, deux se distinguent: le nombre d'appels au service client et la consommation en minutes durant la journée. Les clients multipliant les appels au service client au-delà de 4 ou consommant beaucoup de minutes en journée présentent une probabilité considéable de résilier par rapport aux autres.

À l'inverse, la messagerie vocale apparaît comme un facteur protecteur, tandis que la souscription au plan international constitue un facteur de risque fort, probablement lié aux coûts ou à la qualité de ce service.

Du point de vue de la modélisation, la régression logistique et l'arbre de décision identifient bien les clients fidèles, mais peinent à détecter efficacement les churners. Le recours à des méthodes d'amélioration (rééquilibrage des classes, réglage des hyperparamètres) a permis d'augmenter considérablement le rappel pour les clients à risque. Enfin, le modèle Random Forest a obtenu les meilleures performances, avec une précision globale de 93 %, une excellente identification des clients fidèles (96 %) et une détection correcte des churners (81 %). Ce modèle constitue donc une base robuste pour orienter une stratégie de fidélisation ciblée.

Fort de ces résultats, on recommande de :

- Renforcer la gestion du service client et réduire l'insatisfaction ;
  - ✓ Améliorer la résolution au premier appel et réduire le nombre de réclamations répétées ;
  - ✓ Mettre en place un suivi automatique des clients ayant contacté plusieurs fois le service client afin de leur offrir une prise en charge proactive ;
- Réviser et adapter l'offre internationale :
  - ✓ Réévaluer la tarification et la qualité du plan international, identifié comme facteur de risque ;
  - ✓ Développer des offres plus compétitives et modulables pour répondre aux besoins spécifiques des clients internationaux.

- Valoriser les services fidélisants :
  - ✓ Promouvoir la messagerie vocale auprès des clients qui n'y ont pas encore souscrit ;
  - ✓ Développer des fonctionnalités supplémentaires de ce service afin d'en accroître l'attractivité et d'encourager son adoption.
  - ✓
- Mettre en œuvre une stratégie de fidélisation pilotée par l'analytique prédictive :
  - ✓ Intégrer le modèle Random Forest dans les outils opérationnels de l'entreprise pour identifier en temps réel les clients à risque.

- ✓ Déployer des actions ciblées (réductions, offres personnalisées, programmes de fidélité) afin de retenir les clients avant qu'ils ne résilient.

--