République du Sénégal

*****

Un peuple-Un but -Une foi

Agence nationale de la Statistique et de la Démographie

Ecole nationale de la Statistique et de l'Analyse économique

HACKATON BSA 2024-2025

Prédiction du churn client dans le domaine des télécommunications

Par :

SANDJO Larry Shuman

KENNE YONTA Lesline Meralda

MATANG KUETE Josette Victoire

Élèves Ingénieurs statisticiens économistes, 2ème année

Avant-propos

Dans un contexte économique de plus en plus concurrentiel, la fidélisation des clients représente un enjeu stratégique majeur pour les entreprises. C'est dans cette optique que s'inscrit le présent projet, réalisé dans le cadre d'un hackathon organisé le Bureau des Statistiques de l'Amicale de l'ENSAE de Dakar pour le compte de l'année académique 2024-2025.

Ce rapport retrace les différentes étapes de notre démarche : de l'exploration des données jusqu'à la mise en place d'un modèle prédictif et d'un tableau de bord. Il vise à montrer non seulement la pertinence de l'approche analytique adoptée, mais aussi les choix méthodologiques effectués dans un cadre à la fois contraint par le temps et stimulant sur le plan technique.

Ce projet fut aussi l'occasion de travailler en équipe, d'apprendre à prendre des décisions rapidement et de mettre en œuvre des outils de Machine Learning sur un sujet à fort impact business. Nous espérons que ce travail reflète l'engagement et l'intérêt que nous avons portés à cette problématique.

Sommaire

Liste des tableaux et figures

Liste des tableaux

Liste des figures

Présentation des membres de l'équipe

Ce travail a été réalisé par :

- **SANDJO Larry Shuman :** actuellement élève ingénieur statisticien économiste en 2ème année (ISE 2), il intègre l'ENSAE de Dakar pour le compte de l'année académique 2023-2024 en 1ère année de la filière ISE (ISE 1 Math). Dans ce projet, son travail a essentiellement porté sur la modélisation et les prédictions.
- **KENNE YONTA Lesline Meralda :** actuellement élève ingénieure statisticienne économiste en 2ème année (ISE 2), elle intègre l'ENSAE de Dakar pour le compte de l'année académique 2023-2024 en 1ère année de la filière ISE (ISE 1 Math). Dans ce projet, son travail a essentiellement porté sur la conception de l'application sous Streamlit et les recommandations.
- **MATANG KUETE Josette Victoire :** actuellement élève ingénieure statisticienne économiste en 2ème année, elle intègre l'ENSAE de Dakar pour le compte de l'année académique 2023-2024 en 1ère année de la filière ISE (ISE 1 Math). Le long de ce projet, son travail a consisté en l'ingénierie des données via la conception du pipeline.

Lien vers le repositoire Github : LarrySANDJO/Telecom_churn: Project to predict churn

Lien vers l'application Streamlit : Dashboard Churn Télécom · Streamlit

Lien vers le dataset : Data Science Nigeria Telecoms Churn | Kaggle

Introduction

1. Contexte et justification

L'adage ***« Un tiens vaut mieux que deux tu l'auras »*** traduit une réalité concrète dans le secteur des télécommunications. Pourquoi courir après de nouveaux clients quand ceux que l'on a risquent de partir ? La fidélisation et la rétention client sont devenues aujourd'hui très importantes avec l'accroissement de la concurrence et la diversité des offres sur le marché.

Face à une telle compétitivité, une stratégie de marketing défensive revêt beaucoup d'importance. **Au lieu de tenter d'acquérir de nouveaux clients ou d'attirer les abonnés loin de la concurrence, le marketing défensif s'intéresse plutôt à la réduction des départs de ces clients** selon (Ahn et autres 2006), surtout qu'il est cinq fois plus coûteux d'acquérir un nouveau client que d'en garder un selon (Frederick et Reichheld 1996).

Un client « Churn » dans le secteur des télécommunications fait référence à un client qui cesse sa relation avec l'entreprise et il est probable que ce client rejoindra une entreprise concurrente. Anticiper le churn devient ainsi crucial pour adapter les stratégies commerciales, orienter les campagnes marketing et maintenir une base clientèle solide. **Dans ce contexte, la donnée devient une alliée précieuse.** Les données, nombreuses et variées (consommation, type d'abonnement, …), constituent une mine d'or dans cette quête.

2. Problématique

Dans un secteur aussi concurrentiel que celui des télécommunications, la volatilité des clients pousse les opérateurs à miser sur la fidélisation proactive. **Mais encore faut-il savoir qui risque de partir, quand, et pourquoi**. Nous disposons d'un jeu de données du hackathon de **Data Science Nigeria en 2018 (***Data Science Nigeria Telecoms Churn | Kaggle***)** portant sur les informations de consommateurs d'une entreprise locale. Notre problématique se décline comme suit : peut-on, grâce aux données disponibles, prédire de manière fiable le départ potentiel d'un client ?

3. Objectifs

Notre projet s'inscrit dans le domaine de la Business Intelligence. Il vise à prévoir le risque de churn à partir des données de consommation des clients. L'objectif ultime étant de **fournir à l'opérateur des leviers d'action concrets pour retenir ses clients**, en exploitant les informations disponibles. Plus spécifiquement, il s'agit de :

- Concevoir une solution de machine learning capable de prédire le churn à partir de données clients préalablement traitées ;

- Fournir un outil d'aide à la décision utilisable par les équipes métier via une interface intuitive.

4. Méthodologie

Partant de notre base de données, notre démarche s'articulera autour de cinq étapes : l'exploration puis le prétraitement des données, la modélisation (tests de plusieurs algorithmes supervisés) et l'évaluation des performances à l'aide de métriques adaptées (précision, F1-score, ROC AUC) sous Python ainsi que la conception d'une application Streamlit pour visuChaque étape a été guidée par l'objectif d'optimiser la prédiction du churn tout en assurant l'interprétabilité des résultats.

5. Plan

Le présent rapport est subdivisé en quatre parties. Tout d'abord une brève *revue de la littérature sur les déterminants du churn*, ensuite *l'analyse exploratoire et l'ingénierie des données.* S'en suit la *modélisation* avant la présentation *de l'application* et des différentes recommandations.

Revue de la littérature sur les déterminants du churn

Par définition, le mot *churn* (traduit en français par **attrition**) est né de la contraction en anglais des mots *change* et *turn*. Il décrit le phénomène de perte d'un client. Il est mesuré par le taux de churn qui représente le pourcentage de clients perdus sur une période donnée par rapport au nombre total de clients au début de cette période.

Selon (Tsai, Lu, 2009), les clients Churn du secteur des télécommunications peuvent être classés en deux catégories principales : **involontaire et volontaire**.

- Les clients Churn involontaires : sont les abonnés que l'opérateur de télécommunication décide de supprimer pour de multiples raisons telles que la fraude, le non-paiement… etc.
- Les clients Churn volontaires : peuvent être décrit comme la fin du service par l'abonné.

La prévision du phénomène de Churn repose essentiellement sur des données structurées, car elles sont plus accessibles et plus faciles à exploiter. Le tableau suivant résume les indicateurs du Churn dans les télécommunications.

Tableau 1 : Indicateurs utilisés pour la prévision du Churn dans le domaine des télécommunications

Catégorie de l'indicateur	Indicateurs
Indicateurs d'usage ou de trafic	Minutes d'utilisation (MOU) : nombre total de minutes d'appels sortants effectués par l'abonné durant une période déterminée Fréquence d'utilisation (FOU) : nombre total d'appels sortants sont effectués par l'abonné durant une période déterminée Sphère d'influence (SOI) : capte la puissance de l'influence de l'abonné et est défini comme le nombre total de récepteurs distincts contactés par l'abonné Average Revenue Per User (ARPU) : chiffre d'affaires mensuel moyen réalisé par client (Revenu par ligne d'abonné mobile) Durée en minutes d'utilisation lors des périodes de pic Durée en minutes d'utilisation hors périodes de pic Durée en minutes des appels vers l'international Durée en minutes des appels nationaux Nombre d'appels sortants La durée de service (LOS), qui est la différence entre la date d'activation et la date de terminaison du service Nombre moyen de tentatives d'appels Nombre moyen d'appels entrants d'une durée de moins d'une minute La durée des appels Off-Net et On-Net Nombre d'utilisations du Roaming Nombre SMS/MMS envoyé

Indicateurs de valeur	Montant des appels nationaux par minutes hors minutes gratuites Montant total des appels nationaux et internationaux Montant total égal à la somme du coût total des appels plus le coût de la tarification Montant moyen d'une minute, tarifs et appels internationaux inclus Valeur actuelle nette (VAN) : le calcul de la VAN intègre le revenu brut (somme totale des recharges et des appels entrants), les coûts de gestion et le coût d'acquisition. Valeur à terme des clients (LTV) : est définie par le revenu total généré par un client tout au long de sa vie de client Montant du Roaming
Indicateurs relatifs au client	Type d'appel (mobile-mobile, fixe-mobile, etc.) Nombre de blocages ou de suspensions de service par l'opérateur (dus à un retard de paiement de facture par exemple) Nombre de services optionnels utilisés Statut de la ligne du client (actif, passif, suspendu) Nombre de factures impayées Nombre total des différents plans tarifaires par lesquels le client est passé Plan tarifaire actuel du client
Indicateurs démographiques	Statut matrimonial, Âge du client, Catégorie socioprofessionnelle, Typologie du client (entreprise, particulier, etc.)
Indicateurs d'interaction avec l'opérateur	Moyenne des minutes d'appels aux centres d'appels Nombre de consultations de crédit Nombre de réclamations Temps de résolution des problèmes Nombre de points de fidélité cumulés
Indicateurs de qualité de service	Nombre d'appels perdus, Nombre moyen d'appels terminés correctement

Source : adapté à partir (WARDY et BERRADA, 2012, P08).

Le tableau nous a fourni une synthèse des indicateurs utilisés dans l'industrie des télécommunications pour prévenir le churn afin d'y faire face et d'éviter de perdre les clients. Chaque entreprise a donc le choix de choisir les indicateurs les plus performants et les plus adéquats avec son activité.

Exploration et ingénierie des données

Notre travail repose sur un jeu de données du hackathon de **Data Science Nigeria en 2018.** Il comporte un échantillon de 1400 clients pour lesquels 15 variables ont été collectées (hormis la variable d'identification), couvrant quatre des six catégories d'indicateurs mentionnés dans la revue de la littérature. Cette partie vise à **mieux comprendre la structure, les tendances et les particularités** du jeu de données et à faire les imputations nécessaires.

1. Description des variables

D'emblée, le tableau suivant donne la description des différentes variables.

Tableau 2 : Description des variables du jeu de données

Variables	Description	Nature	Indicateur
Customer ID	Identifiant client	Textuelle	
network_age	Ancienneté du client sur le réseau (en jours)	Quantitative	Usage
Customer tenure in month	Ancienneté du client sur le réseau (en mois)	Quantitative	Usage
Total Spend in Months 1 and 2 of 2017	Dépenses totales du client durant les mois de juillet et août 2017	Quantitative	Valeur
Total SMS Spend	Montant total dépensé en SMS par un client	Quantitative	Valeur
Total Data Spend	Dépenses totales en données par un client	Quantitative	Valeur
Total Data Consumption	Volume total de données consommées par un abonné, en kilo-octets (KB), sur la période étudiée.	Quantitative	Valeur
Total Unique Calls	Nombre total d'appels uniques passés par un abonné durant la période étudiée.	Quantitative	Usage
Total Onnet spend	Dépenses totales en appels vers des abonnés du même opérateur	Quantitative	Valeur
Total Offnet spend	Dépenses totales en appels vers d'autres opérateurs	Quantitative	Valeur
Total Call centre complaint calls	Nombre total d'appels de réclamation au service client	Quantitative	Interaction
Network type subscription in Month 1	Type d'abonnement réseau au mois 1 (juillet), pouvant refléter son type d'appareil (2G, 3G ou autre).	Catégorielle	Relatif au client
Network type subscription in Month 2	Type d'abonnement réseau au mois 2 (août) pouvant être 2G, 3G ou autre	Catégorielle	Relatif au client
Most Loved Competitor network in in Month 1	Réseau concurrent préféré au mois 1, pouvant donner une indication sur l'opérateur vers lequel il pourrait migrer.	Catégorielle	Usage
Most Loved Competitor network in in Month 2	Réseau concurrent préféré au mois 2	Catégorielle	Usage
Churn Status	Statut de churn : 1 si le client s'est désabonné et 0 sinon	Quantitative	

Source : Auteurs

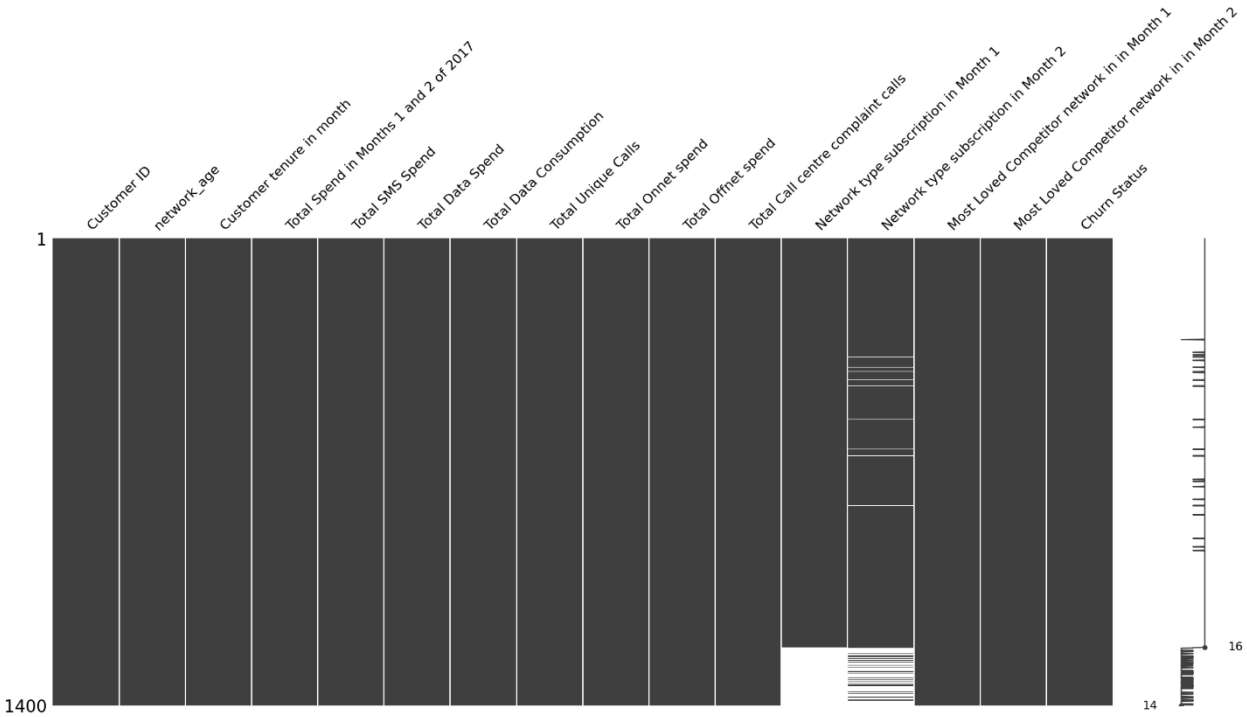
Ainsi, le dataset comporte une variable textuelle, 10 quantitatives et 5 qualitatives.

2. Traitement des données

D'entrée de jeu, les faits suivants ont été relevés dans notre base de données :

- network_age comporte 3 valeurs négatives ce qui est aberrant. Customer tenure in month est donc également affecté (car Customer tenure in month = network_age /30) ;
- Les variables quantitatives comportent chacune pléthore de valeurs extrêmes ;
- **321 valeurs manquantes** ont été détectées soit 1,4% du total et uniquement sur les variables qualitatives : Network type subscription in Month 1 et Network type subscription in Month 2 comportent respectivement 175 (12,5%) et 144 (10,3%) valeurs manquantes ; les variables Most Loved Competitor network in in Month 1 et Most Loved Competitor network in in Month 2 quant à elles ont une valeur manquante chacune.

Figure 1 : Graphique des valeurs manquantes



Source : Calcul des auteurs

Ceci étant, dans le pipeline, nous avons supprimé les individus dont l'ancienneté sur le réseau est négative. Par ailleurs, les informations manquantes sur le concurrent préféré des clients aux mois 1 et 2 ont été imputées par le mode.

En ce qui concerne les variables *Network type subscription in Month 1* et *Network type subscription in Month 2*, nous avons créé la variable ***Networkupgrade*** pour identifier les individus qui sont passés de la 2G à la 3G, de la 3G à la 2G ou sont restés sur le même type de réseau. Il en ressorti que plus de 80% des individus sont restés sur le même réseau. Ainsi, nous avons imputé de la manière suivante :

- Si pour un individu, le réseau auquel il a souscrit au mois 1 est renseigné et pas celui du mois 2, nous attribuons à *Network type subscription in Month 2* la valeur de *Network type subscription in Month 1,* et vice-versa car peu d'individus ont changé de réseau.
- Par suite, nous avons imputé le reste des données manquantes par le mode, pour chacune des deux variables (cas où ni *Network type subscription in Month 1* ni *Network type subscription in Month 2* n'est renseigné).

Nous avons par ailleurs créé dans le pipeline les variables Total SMS Spend_ratio, Total Data Spend_ratio, Total Onnet spend_ratio ,Total Offnet spend_ratio, rendant compte de la part des dépenses totales en SMS (resp. data, appels dans le réseau et appels hors réseau) dans les dépenses totales du mois 1 et 2. Les données aberrantes ont par la suite été imputées en utilisant la méthode de l'intervalle interquartile (IQR) : borne inférieure $= Q1 - 1.5 \times IQR$ et borne supérieure $= Q3 + 1.5 \times IQR$.

Il est également important de noter qu'au cours de la phase de préparation des données, les différents types de données statistiques auxquels nous avons été confrontées ont nécessité des traitements spécifiques. Plus précisément, nous avons appliqué la normalisation aux variables quantitatives car leurs distributions étaient asymétriques et le *frequency encoding* aux variables qualitatives car présentant beaucoup de modalités.

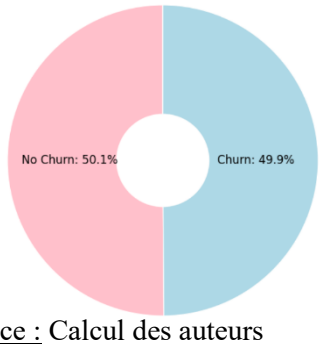
3. Statistiques descriptives univariées

3.1. Distribution de la variable cible

Après nettoyage de la base, notre échantillon comporte 1397 individus dont 50,1% sont encore actifs sur le réseau de de l'entreprise, soit un taux de churn de 49,9%, ce qui est très alarmant pour une entreprise dans un secteur aussi concurrent que les télécommunications.

NB : Nous sommes dans un problème de Machine Learning où la variable cible est équilibrée.

Figure 2 : Fréquence d'attrition des clients



Source : Calcul des auteurs

3.2. Distribution des autres variables

☞ Plus de la moitié des clients (56%) ont souscrit à la 3G au mois de juillet 2017. Ce chiffre a connu une légère hausse en août, passant ainsi à 58%. C'est dire que la majorité des utilisateurs est de cette entreprise est enclin à utiliser la 3G.

☞ *PQza* et *Uxaa* sont les deux concurrents préférés des clients au mois 1 avec 24,4% et 23,1% respectivement d'utilisateurs provenant de l'entreprise. Pour ce mois, il n'y a pas de

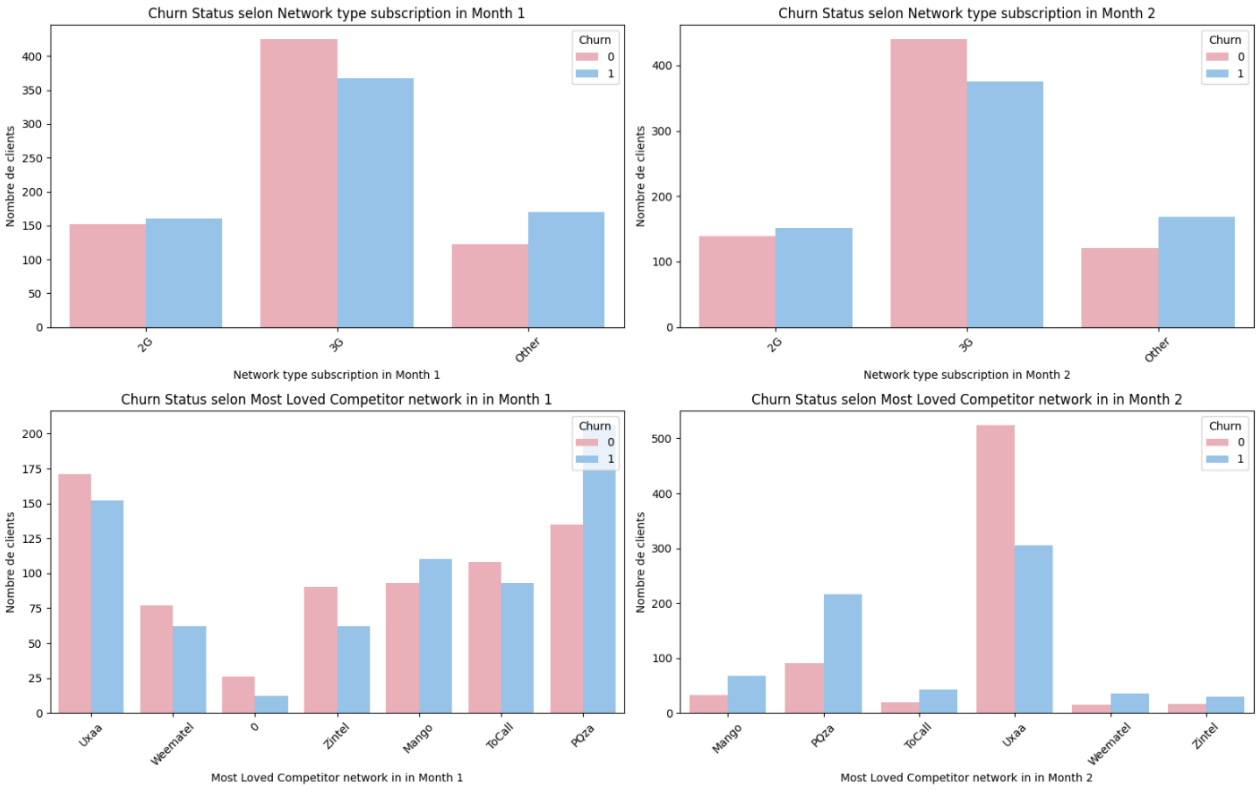
concurrent qui domine largement les autres contrairement au mois 2 où *Uxaa* est largement en tête avec 59,3%, devant *PQZa* (22.0%). Ces résultats révèlent que Uxaa est un sérieux concurrent pour l'entreprise.

☞ Les variables quantitatives quant à elles présentent des distributions asymétriques étalées vers la droite, ainsi que de faibles valeurs pour la médiane. Du fait de l'imputation, il n'y a plus de données aberrantes.

4. Relation entre la variable cible et les autres variables

L'analyse des graphiques ci-dessous indique que les types de réseau influencent légèrement le churn tandis que **les préférences vis-à-vis des concurrents** sont des **indicateurs beaucoup plus forts**. Au mois 1, le churn est **très marqué** pour les clients préférant PQza. Au mois 2, le churn est beaucoup plus important pour ceux préférant Uxaa. Cela confirme une fuite vers certains concurrents, en particulier PQza et Uxaa.

Figure 3 : Répartition de Churn Status selon les variables qualitatives

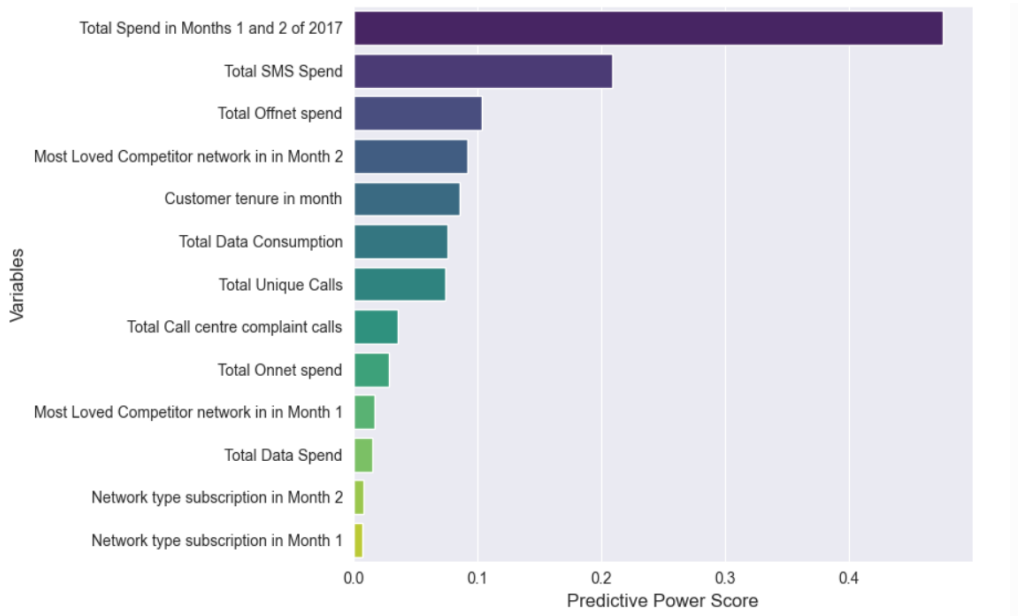


Source : Calcul des auteurs

L'analyse des relations entre la variable de churn et les variables quantitatives révèle que les variables comme *Total Call centre complaint calls, Total Onnet spend, Total Data Spend* ne sont pas déterminantes pour rendre compte du churn des clients. En revanche, les autres variables se révèlent être significativement liées au churn.

Le graphique ci-dessous présente le pouvoir prédictif de nos variables d'origine.

Figure 4 : Top des variables par pouvoir prédictif



Source : calcul des auteurs

5. Corrélation entre les variables

De fortes corrélations ont été observées entre les variables notamment *Network type subscription in Month 1* et *Network type subscription in Month 2* (0.86) ; *Total Unique Calls et Total Spend in Months 1 and 2 of 2017* (0.65) ; *Consistent_competitor* et *Most Loved Competitor network in in Month 1 (0.52).*

Ceci étant, dans le pipeline, nous avons éliminé les variables *Network type subscription in Month 1, Total Unique Calls, Consistent_competitor* ainsi que toutes celles qui ont servi à créer les variables ratios *(Total SMS Spend, Total Data Spend, Total Onnet spend ,Total Offnet spend).*

Méthodologie de Modélisation

1. Choix du Modèle

Pour choisir le modèle de machine Learning le plus performant, plusieurs algorithmes ont été testés, notamment : Régression Logistique, Random Forest, XGBoost, SVM, KNN, Naive Bayes, Perceptron, LightGBM, AdaBoost, …

Le choix du meilleur modèle s'est basé sur trois critères de performance :

- **L'accuracy** : la proportion globale de prédictions correctes.
- **Le F1-Score** : la moyenne harmonique entre la précision et le rappel.
- **L'AUC-ROC** (Area Under the ROC Curve) : mesure la capacité du modèle à distinguer entre les classes positives et négatives.

Le modèle ayant obtenu les meilleures performances est **XGBoost**. 🏆

2. Optimisation des Hyperparamètres

Afin d'améliorer les performances de chaque modèle, une **optimisation des hyperparamètres** a été réalisée en utilisant **GridSearchCV** avec validation croisée. Cette approche systématique permet de tester plusieurs combinaisons de paramètres et de sélectionner celle qui maximise les performances.

Le modèle **XGBoost (eXtreme Gradient Boosting)** s'est démarqué comme étant le plus performant sur notre jeu de données. Il s'agit d'un algorithme d'ensemble basé sur le boosting, qui fonctionne en entraînant successivement des arbres de décision faibles pour corriger les erreurs des modèles précédents.

3. Performances

Le modèle XGBoost, après optimisation des hyperparamètres, a obtenu les meilleurs scores.

Tableau 3 : Meilleurs paramètres du modèles XGBOOST

learning_rate: **0.05**	max_depth: **20**	n_estimators: **100**	subsample: **0.8**

Source : Calcul des auteurs

Tableau 4 : Résultats sur les données de test

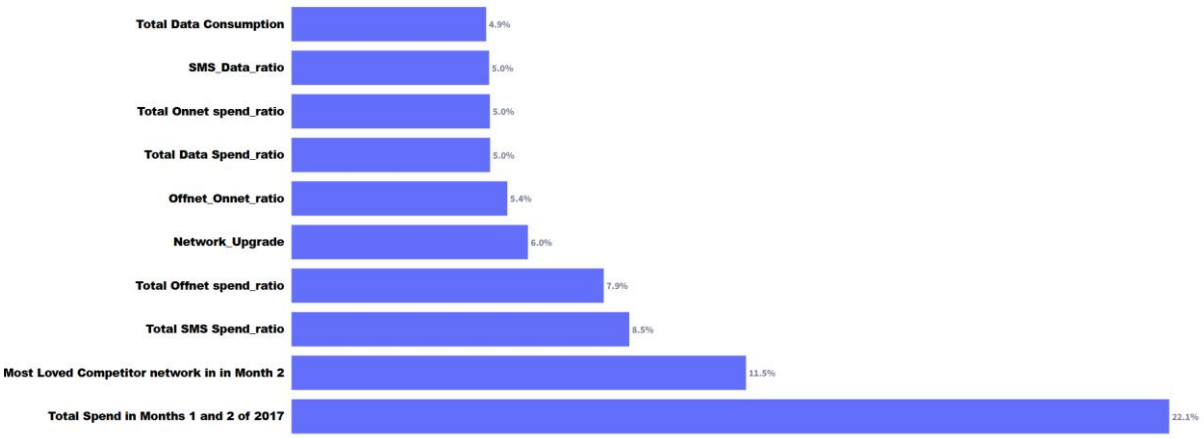
Métrique	Valeur obtenue
Accuracy	**0.8411**
F1-Score	**0.8268**
AUC-ROC	**0.9185**

Source : Calcul des auteurs

Ces résultats indiquent que le modèle est capable de :

- ☞ Bien classifier les clients churn/non-churn (accuracy de 84 %),
- ☞ Maintenir un bon équilibre entre précision et rappel (F1-score de 82 %),
- ☞ Très bien distinguer les clients à risque des autres (AUC-ROC de 91 %).

Figure 5 : Importance des variables dans le modèle XGBOOST



Source : Calcul des auteurs

Présentation du dashboard

Pour une visualisation instantanée des dynamiques réseau et une transformation rapide des données complexes en insights actionnables pour une prise de décision stratégique, le dashboard est l'outil idéal. Un dashboard a donc été développé avec **Streamlit**, un framework Python simple et puissant qui permet de créer des interfaces web interactives pour la data science et le machine learning. Grâce à Streamlit, les utilisateurs peuvent visualiser des données, interagir avec des modèles prédictifs et explorer des résultats en temps réel, le tout sans avoir besoin de compétences en développement web.

Dédié à l'analyse et la prédiction de l'attrition client, ce dashboard s'articule autour de trois pages principales et dispose d'un système de filtrage transversal pour une analyse personnalisée.

1. Système de filtrage global

Dans la barre latérale, nous avons implémenté trois filtres principaux qui affectent dynamiquement l'ensemble des visualisations :

- **Filtre par type de réseau** : permet de sélectionner un ou plusieurs types d'abonnements réseau
- **Filtre par niveau de risque** : segmente les clients selon leur probabilité de churn (Faible, Moyen, Élevé)
- **Filtre par segment client :** Classifie les clients selon leur dépense totale (Basse, Moyenne-Inférieure, Moyen-Supérieure, Élevée)

Afin de d'avoir une idée rapide de l'impact des filtres, un compteur affiche dynamiquement le nombre de clients correspondant aux critères sélectionnés dans ces filtres.

2. Pages de l'application

2.1. Page d'accueil et d'indicateurs de performances (KPIs)

Cette page présente une vue d'ensemble de la situation actuelle avec des **indicateurs clés (KPIs) et des visualisations d'analyse de l'attrition des abonnés** suivant les critères sélectionnés dans le filtre.

KPIs principaux

- ✓ **Le taux de Churn actuel** qui mesure la santé de l'entreprise en fournissant le pourcentage de clients perdus par l'entreprise au cours des 2 mois étudiés.
- ✓ **Taux d'actifs** présente le pourcentage de clients toujours actifs dans la base, indicateur fondamental de rétention client.

- ✓ **L'ancienneté moyenne** qui mesure la fidélité moyenne du client afin de contextualiser le churn.
- ✓ **Le taux de plaintes** mesure de satisfaction client qui est souvent précurseur du churn.

Visualisations analytiques

- ✓ **Segmentation Client :** présentée au travers d'un secteur circulaire, elle illustre les pourcentages des groupes de clients selon la valeur qu'ils apportent à l'entreprise afin d'identifier les groupes majoritaires. Trois segments distincts ont été créés selon le niveau de dépense sur les 2 premiers mois de 2017 en utilisant le **clustering**:
  - *Le segment de dépense "**Faible**" (dépenses < 2 332,63) : clients peu engageants*
  - *Le segment de dépense **"Moyenne"** (2 332,63 ≤ dépenses < 9 969,21) : clients modérément actifs*
  - *Le segment de dépense **"Élevé"** (dépenses ≥ 9 969,21) : clients hautement rentables*
- ✓ **Churn par segment :** il s'agit d'un diagramme à barres qui présente les pourcentages de churn dans chaque segment suscité, afin d'identifier les segments les plus vulnérables, guidant ainsi les efforts de rétention.
- ✓ **Churn et réclamations :** il s'agit d'un diagramme en barres groupées qui a pour but de présenter la corrélation entre le nombre de plaintes et le risque de churn. Les clients ont été segmenté en classe selon leur niveau de plaintes (0 : aucune plainte, 1-3 ans : faible plainte, pour plus de trois plaintes, le niveau est considéré comme élevé). S'il montre par exemple que le taux de churn est plus grand pour les clients de niveau de plaintes élevé, l'entreprise devra mettre un accent particulier à la résolution des plaintes.
- ✓ **Churn et ancienneté :** ce diagramme en barres groupées a pour but de présenter la relation entre l'ancienneté des clients et leur risque d'attrition. Les clients ont été segmentés en classes selon leur durée d'ancienneté (≤12 mois : Nouveau, ≤36 mois : Etabli, >36 mois : Fidèle). S'il montre par exemple que le taux de churn est plus élevé chez les nouveaux clients, l'entreprise devra concentrer ses efforts de fidélisation sur les 12 premiers mois de la relation client pour réduire le risque de départ.
- ✓ **Concurrents préférés :** ce diagramme identifie vers quels concurrents les clients partent, permettant d'analyser les stratégies de ces concurrents pour mettre en place leurs propres stratégies de rétention clients.
- ✓ **Rentabilité des réseaux** : Ce diagramme en barres permet de visualiser les types de réseaux offerts par l'entreprise suivant leur contribution au chiffre d'affaires de l'entreprise. Un indicateur qui va permettre de savoir quel type de réseau est le plus rentable afin de voir comment inciter les clients vers l'utilisation de ce réseau.

- ✓ **Répartition des dépenses par type de réseau :** diagramme à barres qui a pour but de montrer la contribution de chaque type de réseau au chiffre d'affaires global. Ce graphique utilise le type de réseau utilisé au mois 2 et calcule le pourcentage des dépenses totales pour chaque type. S'il montre qu'un type de réseau génère une part significativement plus importante des revenus, l'entreprise devra prioriser les investissements dans ce réseau pour maximiser sa rentabilité.
- ✓ **Répartition des dépenses par service :** cette représentation circulaire décompose le panier moyen des clients entre différents services **(SMS, Data**, appels **Onnet/Offnet)**. La visualisation des préférences de consommation éclaire la direction produit sur les services à privilégier. Une prédominance des dépenses en data, par exemple, orienterait naturellement les efforts marketing vers des forfaits enrichis en données mobiles.
- ✓ **Contribution des segments aux dépenses totales :** graphique à barres qui quantifie la valeur économique de chaque segment client. En classant les segments par leur contribution aux revenus, il permet d'optimiser l'allocation des ressources marketing. L'identification des segments premium justifie le développement de programmes de fidélisation sur mesure, tandis que les segments moins rentables peuvent faire l'objet de stratégies de développement ciblées.
- ✓ **Migrations réseaux :** Cette visualisation dynamique montre les migrations d'abonnements réseau (2G, 3G, Other) entre deux mois consécutifs via une matrice de flux (Sankey ou heatmap). Elle permet d'identifier :
  - Les **tendances de mise à niveau** vers des réseaux plus performants : une migration massive vers "Other" depuis 2G/3G peut indiquer un désintérêt pour les réseaux traditionnels, créant une opportunité de promotion 4G/5G.
  - Les **risques de rétrogradation** pouvant signaler des problèmes techniques ou tarifaires : un flux significatif 3G → 2G signale potentiellement une dégradation de service ou une tarification inadaptée. Il serait donc intéressant dans ce cas de lancer des campagnes de fidélisation ciblées pour ces clients en migration à risque.

2.2. Page de prédictions

Cette page se concentre sur les prévisions de churn et présente des indicateurs de prise de décision basés sur notre modèle d'apprentissage automatique. Ainsi, nous y avons :

KPIs principaux

Nombre de clients prédits à haut risque : le risque est considéré comme haut ici lorsqu'il dépasse 60%. Il s'agit là du taux clients nécessitant une attention immédiate.

Visualisations analytiques

- ✓ **Répartition des clients à haut risque par segment :** ce graphique permet de voir les clients les plus à risque de l'entreprise, la part de chaque segment.
- ✓ **Taux de clients prédits à haut risque par segment :** Ce graphique permet de voir dans chaque segment, la proportion des clients à haut risque afin de savoir dans quel segment le risque de churn prévaut le plus et de mettre en place des stratégies de rétention les ciblées.
- ✓ **Analyse des facteurs d'influence :** cette section présente les variables qui influencent le plus la décision d'attrition d'un client, permettant ainsi aux équipes marketing et service client de cibler leurs actions sur les facteurs les plus impactants.

Prédiction proprement dite

Cette page offre enfin la possibilité de prédire à partir des caractéristiques de d'un ou plusieurs clients, leur probabilité de quitter l'entreprise. Elle permet ainsi de prendre assez rapidement des décisions stratégiques pour anticiper et éviter le départ de clients.

2.3. Page des recommandations

Cette page offre une vision stratégique des clients nécessitant une attention particulière, permettant d'orienter efficacement les actions commerciales et de fidélisation.

Eléments clés mis en évidence

Les éléments clés mis en évidence dans cette page sont les suivants :

✓ **Les clients à risque élevé de churn :** cette liste identifie les abonnés les plus susceptibles de quitter le réseau, permettant d'intervenir proactivement avec des offres ciblées avant leur départ. L'identification précoce de ces clients peut significativement réduire le taux d'attrition et préserver le revenu récurrent du réseau.

✓ **Les clients les plus anciens :** ces fidèles de longue date représentent un capital confiance précieuse pour la marque. La reconnaissance de leur loyauté par des promotions dédiées renforce non seulement leur satisfaction mais améliore également l'image de l'entreprise auprès de l'ensemble de la clientèle.

✓ **Les clients à forte valeur :** ce segment contribue de manière disproportionnée au chiffre d'affaires. Les offres premium qui leur sont destinées augmentent leur sentiment d'exclusivité tout en maximisant leur valeur vie client (LTV), créant ainsi un cercle vertueux d'engagement.

✓ **Les clients les plus récents :** l'analyse de ces nouveaux venus permet d'évaluer l'efficacité des dernières campagnes d'acquisition et d'affiner le processus d'intégration pour optimiser leur expérience initiale.

✓ **Clients avec ratio Offnet/Onnet > 3 :** ces utilisateurs communiquent majoritairement hors du réseau, générant des coûts d'interconnexion élevés. Des offres adaptées peuvent rééquilibrer leur profil d'utilisation et améliorer la marge opérationnelle.

✓ **Clients ayant ratio Offnet/Onnet < 1 :** ces abonnés privilégient les communications au sein du réseau, représentant un profil économiquement avantageux. Leur fidélisation contribue à la rentabilité globale de la base client. Il faut donc multiplier les offres intra-réseau pour ces clients.

✓ **Clients ayant Migration à Risque :** Ces clients, qui rétrogradent volontairement ou non de la 3G vers la 2G, représentent un profil à risque pour l'entreprise. Ils sont susceptibles de quitter l'opérateur si leurs besoins ne sont pas comblés. Pour comprendre les raisons de leurs rétrogradations, des enquêtes ciblées sont nécessaires, tandis que des stratégies de rétention comme les bonus data gratuits et l'amélioration de la couverture 3G dans leurs zones géographiques peuvent prévenir ces migrations indésirables.

Quelques croisements stratégiques à explorer

Le système de filtrage croisé, présenté plus haut, offre une granularité d'analyse exceptionnelle. Il permet par exemple sur cette page de recommandations, d'explorer les croisements suivants :

- ✓ Identification des **clients à forte valeur présentant un risque de churn élevé** pour des interventions commerciales urgentes et personnalisées.
- ✓ Ciblage des **clients anciens utilisant encore la 2G** pour des migrations technologiques incitatives vers des services plus performants et rentables.
- ✓ Analyse des **nouveaux clients avec un ratio Offnet/Onnet déséquilibré** pour ajuster rapidement la communication d'accueil et les offres de bienvenue.
- ✓ Focalisation sur les **clients à dépense moyenne montrant des signes précoces d'attrition** pour maximiser le retour sur investissement commercial.

Cette vision multidimensionnelle de la base client transforme les données en actions commerciales précises et mesurables, optimisant simultanément satisfaction client et performance économique.

Conclusion

Ce projet a permis de mettre en œuvre une démarche complète de data science dans le but de prédire le churn client à partir de données réelles. Après avoir exploré et préparé les données, plusieurs modèles de machine learning ont été testés et comparés selon des métriques de performance pertinentes.

Le modèle **XGBoost**, optimisé via GridSearchCV, s'est révélé le plus performant, atteignant une **accuracy de 84,1 %**, un **F1-score de 82,7 %** et un **AUC-ROC de 91,8 %**. Ces résultats témoignent d'une bonne capacité de prédiction et d'un bon équilibre entre les différentes métriques, en particulier pour identifier les clients à risque de départ.

Nous avons également développé un **dashboard interactif** sur Streamlit, permettant de visualiser les résultats et de faciliter la prise de décision. Enfin, en nous appuyant sur les facteurs influençant le churn, nous avons formulé des recommandations concrètes pour aider l'entreprise à mieux cibler ses actions de fidélisation.

Ce travail ouvre la voie à une application concrète d'aide à la décision pour les opérateurs, notamment en Afrique, en facilitant la mise en place d'actions ciblées de fidélisation.

Bibliographie

[1] Ahn, J. H., Han, S. P., & Lee, Y. S. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications policy*, *30*(10-11), 552-568.

[2] Diaw, M. (2024) Cours machine de learning 1 : modèles de classification.

[3] Lanseur, A., & Ait Sidhoum, H. (2022). Les déterminants du churn client dans le secteur des télécommunications : étude des trois opérateurs de la téléphonie mobile en Algérie. Université de Bejaia.

[3] Reichheld, F. F., & Teal, T. (1996). The loyalty effect: The hidden force behind growth, profits and lasting. *Harvard Business School Publications, Boston*.

[4] WARDY, S., & BERRADA, I (2012). Double problematique du churn et du turnover pour les organisations : definitions et etat de l'art. *Microsoft Word - WARDY-BERRADA.doc*