جامعة مولاي إسماعيل
UNIVERSITÉ MOULAY ISMAÏL

المدرسة الوطنية العليا للفنون و المهن
ÉCOLE NATIONALE SUPÉRIEURE D'ARTS ET MÉTIERS

Prédiction de l'Attrition des Clients Télécom

(Market and Customer Analysis)

Filière : Génie industriel intelligence artificielle et DATA science

Encadré par : Monsieur Hosni

Réalisé par :

- HJIRT Soufia
- ISSOUMMOUR Amal
- DRIEF Nisrine
- BOUTGHART Alae
- AOUSSAR Wissal
- EL RHAZI Kaoutar

Le 13 Mai 2025

Table des matières

1 Introduction

L'objectif de ce projet est d'analyser les données de clients d'une entreprise de télécommunications pour prédire l'attrition (départ) des clients. Ce problème est crucial pour les opérateurs télécoms car il est généralement plus coûteux d'acquérir de nouveaux clients que de retenir les clients existants.

Dans un secteur aussi concurrentiel que les télécommunications, la rétention des clients est devenue un enjeu stratégique majeur. La capacité à identifier de façon précoce les clients susceptibles de résilier leur contrat (churn) permettrait aux opérateurs de mettre en place des actions ciblées pour les fidéliser avant qu'ils ne partent vers la concurrence.

1.1 Objectifs du projet

Ce projet vise deux objectifs principaux :

— Développer un modèle prédictif capable d'identifier avec précision les clients à risque d'attrition

— Déterminer les facteurs clés qui influencent la décision d'un client de quitter l'opé- rateur

1.2 Méthodologie

Pour atteindre ces objectifs, nous suivrons une méthodologie complète de data mining comprenant :

1. Chargement et inspection des données
2. Nettoyage et prétraitement des données
3. Analyse exploratoire des données
4. Feature engineering
5. Modélisation par machine learning
6. Évaluation des modèles
7. Optimisation des hyperparamètres
8. Interprétation des résultats

Chaque étape sera documentée en détail dans ce rapport, avec une justification des choix techniques effectués et une analyse critique des résultats obtenus.

2 Description des données

2.1 Source et contenu

Les données utilisées dans cette étude proviennent du jeu de données "WA_Fn-UseC_- Telco-Customer-Churn.csv" disponible sur Kaggle. Ce dataset contient 7043 entrées avec 21 variables réparties comme suit :

— **Variables démographiques** : genre, âge, présence d'enfants/partenaire, statut senior

— **Services souscrits** : téléphonie, internet, sécurité en ligne, sauvegarde, protection d'appareil, support technique, streaming TV/films

— **Informations contractuelles** : type de contrat, durée d'abonnement, facturation électronique/papier
— **Données financières** : charges mensuelles, charges totales
— **Variable cible** : "Churn" (attrition) - binaire (Oui/Non)

2.2 Pertinence pour la problématique

Ce jeu de données est parfaitement adapté à notre problématique car :

— Il fournit une variable cible explicite (Churn)
— Il couvre une large gamme de facteurs potentiellement influents sur la décision de résiliation
— Il présente une taille d'échantillon suffisante pour l'entraînement de modèles robustes
— Il reflète la réalité commerciale des opérateurs télécoms avec un déséquilibre naturel entre clients fidèles et partants

3 Préparation de l'environnement

3.1 Outils et bibliothèques

Pour ce projet, nous avons utilisé l'écosystème Python qui offre une grande flexibilité et de nombreuses bibliothèques spécialisées pour le data mining :

— **Pandas** et **NumPy** pour la manipulation des données
— **Matplotlib** et **Seaborn** pour la visualisation
— **Scikit-learn** pour les algorithmes de machine learning et l'évaluation des modèles
— **Imbalanced-learn** pour gérer le déséquilibre des classes
— **XGBoost** pour les algorithmes de boosting avancés

Le choix de ces bibliothèques se justifie par leur maturité, leur fiabilité et leur capacité à traiter efficacement les problématiques de classification binaire comme celle de l'attrition client.

4 Chargement et inspection des données

4.1 Méthodes utilisées

Pour cette première étape, nous avons utilisé les fonctions de base de Pandas pour charger et explorer le dataset :

— pd.read_csv() pour importer les données
— data.head() pour visualiser les premières lignes du dataset
— data.info() pour obtenir les types de données et identifier les valeurs manquantes
— data.describe() pour calculer les statistiques descriptives des variables numé- riques

4.2 Observations initiales

Cette inspection préliminaire nous a permis de faire plusieurs constats importants :

— Le dataset contient 7043 observations et 21 colonnes

— Les variables sont mixtes (numériques et catégorielles)

— La colonne "TotalCharges" contient des valeurs manquantes

— La variable cible "Churn" présente un déséquilibre, avec environ 73% de clients fidèles contre 27% de clients ayant résilié

Ces observations initiales ont guidé notre démarche de prétraitement et d'analyse exploratoire dans les étapes suivantes.

5 Nettoyage et prétraitement des données

5.1 Gestion des valeurs manquantes

Notre analyse a révélé la présence de valeurs manquantes dans la colonne "Total-Charges". Nous avons opté pour une approche pragmatique :

1. Conversion de la colonne en format numérique avec pd.to_numeric()
2. Remplacement des valeurs manquantes par la médiane de la colonne

Justification : La médiane a été préférée à la moyenne car elle est moins sensible aux valeurs extrêmes. Cette solution est adaptée car le nombre de valeurs manquantes était limité et n'affecte pas significativement la distribution des données.

5.2 Encodage des variables catégorielles

Les algorithmes de machine learning nécessitant des entrées numériques, nous avons procédé à l'encodage des variables catégorielles :

— Utilisation de LabelEncoder pour convertir les catégories en valeurs numériques

— Application à toutes les colonnes catégorielles sauf 'customerID' qui sert uniquement d'identifiant

Justification : Le LabelEncoder a été choisi pour sa simplicité et son efficacité. Pour les variables binaires comme "gender" ou "Partner", cette approche est parfaitement adaptée. Pour les variables multi-catégorielles comme "Contract", cette méthode préserve l'information catégorielle tout en la rendant exploitable par les algorithmes.

5.3 Séparation des données

Conformément aux bonnes pratiques en machine learning, nous avons divisé notre dataset :

— 80% pour l'ensemble d'entraînement

— 20% pour l'ensemble de test

Justification : Cette répartition couramment utilisée offre un bon compromis entre la taille de l'ensemble d'entraînement (suffisamment grande pour permettre un apprentissage efficace) et celle de l'ensemble de test (suffisamment représentative pour une évaluation fiable).

6 Analyse exploratoire des données (EDA)

6.1 Distribution de la variable cible

Notre première analyse a porté sur la distribution de la variable cible "Churn". Nous avons utilisé la visualisation par sns.countplot() qui a confirmé le déséquilibre signifi- catif entre les classes :

— Environ 73% des clients restent fidèles à l'opérateur

— Environ 27% des clients quittent l'opérateur

Ce déséquilibre, bien que modéré, justifie la mise en place de techniques spécifiques lors de la modélisation pour éviter les biais de prédiction.

6.2 Analyse des corrélations

La matrice de corrélation, visualisée par une heatmap avec sns.heatmap(), a révélé plusieurs relations importantes :

— Forte corrélation négative entre l'ancienneté (tenure) et l'attrition

— Corrélation positive entre les charges mensuelles et l'attrition

— Corrélation entre certains services (comme la fibre optique) et l'attrition

6.3 Insights clés

L'exploration des données a mis en évidence plusieurs facteurs déterminants :

— **Type de contrat** : Les clients avec des contrats mensuels présentent un taux d'at- trition significativement plus élevé que ceux avec des contrats annuels ou biannuels

— **Type de service internet** : Les clients équipés de la fibre optique ont tendance à partir plus souvent que ceux avec une connexion DSL

— **Ancienneté** : Plus un client est ancien, moins il a tendance à résilier son contrat, suggérant une relation de fidélisation progressive

— **Services additionnels** : La présence ou l'absence de certains services comme le support technique ou la sécurité en ligne influence le taux d'attrition

Ces observations préliminaires nous ont guidés dans notre approche de feature engi- neering et de modélisation.

7 Feature Engineering

Pour améliorer la qualité prédictive de nos modèles, nous avons créé de nouvelles variables plus informatives et sélectionné les caractéristiques les plus pertinentes.

7.1 Création de nouvelles variables

Nous avons développé deux nouvelles caractéristiques à forte valeur ajoutée :

7.1.1 Ratio des charges

- **Formule** : $charge\_ratio = \frac{MonthlyCharges}{TotalCharges}$
- **Justification** : Ce ratio permet d'identifier les clients qui paient actuellement un montant mensuel élevé par rapport à leur investissement total, ce qui pourrait in- diquer des augmentations récentes de tarif et donc un risque accru d'attrition

7.1.2 Catégorisation de l'ancienneté

- **Méthode** : Transformation de la variable continue "tenure" en variable catégorielle avec 5 groupes : 0-1 an, 1-2 ans, 2-4 ans, 4-6 ans, 6+ ans
- **Justification** : Cette catégorisation permet de capturer des effets non-linéaires de l'ancienneté sur l'attrition, notamment un risque élevé durant la première année suivi d'une stabilisation

7.2 Sélection des caractéristiques

Pour réduire la dimensionnalité et améliorer la performance des modèles, nous avons appliqué une méthode de sélection statistique :

- **Technique** : SelectKBest avec le test ANOVA F-value (f_classif)
- **Paramètres** : Sélection des 10 meilleures caractéristiques (k=10)
- **Justification** : Cette approche permet de :
  - Réduire le bruit en éliminant les variables peu informatives
  - Diminuer le risque de surapprentissage
  - Améliorer la vitesse d'entraînement des modèles
  - Faciliter l'interprétabilité des résultats

Les caractéristiques sélectionnées incluaient notamment le type de contrat, l'ancien- neté, les charges mensuelles et les principaux services souscrits, confirmant ainsi les ob- servations de notre analyse exploratoire.

8 Modélisation par Machine Learning

8.1 Gestion du déséquilibre des classes

Pour traiter le déséquilibre entre clients fidèles et partants, nous avons appliqué la technique SMOTE (Synthetic Minority Over-sampling Technique) :

- **Principe** : Création d'exemples synthétiques de la classe minoritaire
- **Avantages** :
  - Équilibre des classes dans l'ensemble d'entraînement
  - Amélioration de la détection des clients à risque d'attrition
  - Réduction du biais vers la classe majoritaire
- **Implémentation** : Utilisation de la bibliothèque imbalanced-learn

Cette approche est particulièrement pertinente dans notre contexte où la détection des clients partants (classe minoritaire) présente le plus grand intérêt business.

8.2 Sélection des modèles

Nous avons implémenté et comparé quatre algorithmes de classification :

8.2.1 Régression Logistique

- **Principe** : Modèle linéaire estimant la probabilité d'appartenance à une classe
- **Avantages** : Simple, rapide, hautement interprétable
- **Pertinence** : Utilisé comme baseline pour comparer les modèles plus complexes et identifier les relations linéaires simples

8.2.2 Random Forest

- **Principe** : Ensemble d'arbres de décision dont les prédictions sont agrégées
- **Avantages** : Capture les relations non-linéaires, robuste au bruit, gère naturellement les variables mixtes
- **Pertinence** : Adapté pour modéliser les interactions complexes entre les différents facteurs influençant l'attrition

8.2.3 Support Vector Machine (SVM)

- **Principe** : Recherche d'un hyperplan optimal séparant les classes
- **Avantages** : Efficace en grande dimension, utilise des noyaux pour capturer des frontières non-linéaires
- **Pertinence** : Performant pour trouver des frontières de décision complexes dans l'espace des caractéristiques

8.2.4 XGBoost

- **Principe** : Algorithme de boosting optimisé basé sur les arbres de décision
- **Avantages** : Haute performance, régularisation intégrée, gestion optimale des don- nées déséquilibrées
- **Pertinence** : Excellente capacité à capturer des schémas d'attrition subtils et com- plexes

Notre approche progressive (du plus simple au plus complexe) permet d'évaluer systématiquement le gain en performance apporté par chaque niveau de complexité supplé- mentaire.

9 Évaluation des modèles

Pour évaluer objectivement nos modèles, nous avons utilisé plusieurs métriques complémentaires, particulièrement adaptées aux problèmes de classification binaire déséquili- brée.

9.1 Métriques d'évaluation

— **Précision** : Proportion de prédictions positives correctes parmi toutes les prédic- tions positives
— **Rappel (Recall)** : Proportion de vrais positifs correctement identifiés
— **F1-score** : Moyenne harmonique de la précision et du rappel
— **AUC-ROC** : Mesure de la capacité du modèle à distinguer entre les classes

Dans notre contexte, le rappel est particulièrement important car il représente la capacité à identifier correctement les clients à risque de départ, permettant ainsi des actions de rétention ciblées.

9.2 Résultats comparatifs

Les performances des différents modèles sont résumées dans le tableau suivant :

Modèle	Précision	Recall	F1-score	AUC-ROC
Régression Logistique	0.54	0.78	0.64	0.84
Random Forest	0.57	0.58	0.58	0.82
SVM	0.44	0.63	0.52	0.76
XGBoost	0.56	0.60	0.58	0.82

TABLE 1 – Comparaison des performances des modèles

9.3 Analyse des résultats

— La régression logistique offre le meilleur rappel (0.78) mais au prix d'une précision modérée (0.54)
— Random Forest et XGBoost présentent un meilleur équilibre entre précision et rappel
— Le score AUC-ROC d'environ 0.82-0.84 pour les meilleurs modèles indique une bonne capacité de discrimination
— La précision relativement modeste (0.54-0.57) pour tous les modèles reflète la diffi- culté intrinsèque du problème

Au vu de ces résultats, nous avons sélectionné XGBoost pour l'optimisation des hyperparamètres en raison de son bon équilibre entre précision et rappel, ainsi que pour sa capacité reconnue à s'améliorer significativement après optimisation.

10 Optimisation des hyperparamètres

Pour maximiser la performance de notre modèle XGBoost, nous avons procédé à l'optimisation systématique de ses hyperparamètres.

10.1 Méthodologie

— **Technique** : Recherche par grille (GridSearchCV)
— **Validation** : Validation croisée à 5 plis
— **Métrique d'optimisation** : AUC-ROC

10.2 Paramètres explorés

— n_estimators : [100, 200] - Nombre d'arbres de décision

— max_depth : [3, 5, 7] - Profondeur maximale des arbres

— learning_rate : [0.01, 0.1] - Taux d'apprentissage

Justification : Le choix de ces paramètres et de leurs plages se base sur un com- promis entre exhaustivité de la recherche et temps de calcul. L'AUC-ROC a été préférée comme métrique d'optimisation car elle est moins sensible au déséquilibre des classes que l'accuracy.

10.3 Résultats de l'optimisation

L'optimisation a permis d'améliorer les performances du modèle XGBoost :

— Meilleure configuration : n_estimators=200, max_depth=5, learning_rate=0.01

— Amélioration de l'AUC-ROC : passage de 0.82 à 0.85

— Amélioration significative du rappel sans dégradation excessive de la précision Cette phase d'optimisation démontre l'importance du fine-tuning des hyperparamètres pour maximiser la performance prédictive, particulièrement dans le cas d'algorithmes complexes comme XGBoost.

11 Interprétation des résultats

Au-delà des métriques de performance, l'interprétation des résultats est essentielle pour traduire nos découvertes en insights actionnables pour l'entreprise.

11.1 Importance des caractéristiques

L'analyse de l'importance des caractéristiques dans notre modèle XGBoost optimisé a révélé les facteurs les plus déterminants de l'attrition :

1. **Type de contrat** : Variable la plus influente, confirmant que les clients avec des contrats mensuels sont significativement plus susceptibles de résilier
2. **Ancienneté (tenure)** : Forte relation inverse avec l'attrition, les nouveaux clients étant les plus vulnérables
3. **Charges mensuelles** : Impact positif sur l'attrition, suggérant une sensibilité au prix
4. **Service internet** : Les clients fibres présentent un risque plus élevé
5. **Support technique** : L'absence de ce service est associée à un risque accru

À l'inverse, certaines caractéristiques démographiques comme le genre ou le statut familial se sont révélées peu prédictives, suggérant que l'attrition est davantage liée aux aspects contractuels et de service qu'aux profils démographiques.

11.2 Implications business

Ces résultats permettent de formuler plusieurs recommandations stratégiques :

- **Ciblage prioritaire** : Concentrer les efforts de rétention sur les clients récents avec des contrats mensuels et des charges élevées
- **Approche produit** : Améliorer la qualité et la satisfaction associées aux services fibre optique
- **Services additionnels** : Promouvoir activement le support technique et les options de sécurité qui semblent renforcer la fidélisation
- **Politique tarifaire** : Envisager des offres promotionnelles pour les clients identifiés à risque durant leur première année

La mise en œuvre de ces recommandations, guidée par notre modèle prédictif, pour- rait permettre une réduction significative du taux d'attrition et une amélioration de la rentabilité client à long terme.

12 Conclusion

Ce projet a permis de développer un modèle prédictif performant pour l'identification des clients à risque d'attrition dans une entreprise de télécommunications. En suivant une méthodologie rigoureuse, nous avons :

- Analysé en profondeur les données clients pour comprendre les facteurs d'attrition
- Développé de nouvelles variables explicatives à forte valeur ajoutée
- Comparé différentes approches de modélisation avec une évaluation multidimension- nelle
- Optimisé les paramètres du modèle le plus prometteur
- Extrait des insights actionnables pour guider la stratégie de rétention